AN ASSESSMENT OF DEVIATIONS FROM CONDITIONAL INDEPENDENCE IN BINARY DATA FUSION

Elsabé Smit

Supervisor: Professor J. S. Galpin

School of Statistics and Actuarial Science
University of Witwatersrand

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

Johannesburg, 2011

DECLARATION

I declare that this research report is my own, unaided work. It is being submitted for the Degree of Master of Science to the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

1 March 2011

ABSTRACT

Data fusion is a data integration technique that provides a way to combine information from different sources through a set of common characteristics (variables), thereby creating a single, all-inclusive data source. The success of a fusion largely depends on the accuracy of the underlying assumptions about the relationship between the common variables and the variables unique to each individual data source. The most common model used to fuse data is based on the assumption of conditional independence, which states that the variables unique to each data set (say **Y** and **Z**) are independent given the common variables (say **X**). This analysis evaluates data fusion procedures for binary data under the assumption of conditional independence, and assesses how deviations from this assumption influence the success of the fusion. The degree of conditional independence present in the data is quantified using a function of entropy, namely the conditional mutual information. The impact of the deviation from conditional independence on the success of the fusion is evaluated using the results from a number of different statistical tests, such as the Chi-square goodness-of-fit test and the $\tilde{T}3$-test for a correlation structure, in relation to the level of conditional independence in the data.

Dedicated with all my heart to my father

Awie Smit

1934 – 2007

ACKNOWLEDGEMENTS

This research project would not have been possible without the support and guidance from many people. First and foremost I wish to express my deepest gratitude to my supervisor, Prof. Jacky Galpin, who has supported me throughout this research with patience, guidance and inspiration. I also want to thank all my colleagues at the university for their support and encouragement. The greatest thanks go to my family: my brothers for their love and motivation, and most importantly, my mother for being there for me through the difficult times and sharing the exciting moments, for all the emotional support, for her unconditional love and constant care.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY OF ACRONYMS

AFI	Adjusted Family Income
ARF	Advertising Research Foundation
BARB	Broadcasters Audience Research Board
BEA	Bureau of Economic Analysis
BHPS	British Household Panel Study
CIA	Conditional Independence Assumption
CMI	Conditional Mutual Information
CPS	Current Population Survey
EM	Expectation-Maximization algorithm
FEX	Family Expenditure Survey
IRS	Internal Revenue Service
*k*NN	*k*-Nearest Neighbour
LLR	Local Linear Regression estimator
MAR	Missing At Random
MCAR	Missing Completely At Random
MESP	Measurement of Economic and Social Performance
MLE	Maximum Likelihood Estimates
MNAR	Missing Not At Random
NCVS	National Crime Victimization Survey
NHIS	National Health Interview Survey
NSAF	National Survey of American Families
ODD	Order of Decreasing Difficulty
OTA	Office of Tax Analysis
PIA	Pairwise Independence Assumption
PUS	Public Use Survey
qCIA	Quantified Conditional Independence Assumption measure
RR	Risk Ratio / Relative Risk
SCA	Survey of Consumer Attitudes

SCF	Survey of Consumer Finances
SEO	Survey of Economic Opportunity
SFCC	Survey of Financial Characteristics of Consumers
SNA	United Nations System of National Accounts
SSD	Sum of Squared Deviations
TAM	TV People Meter panels
TGI	Target Group Index
TGR	Target Group Ratings
TM	Tax Model
TF	Tax File

1 INTRODUCTION

1.1 Background

In the market research industry, large amounts of data are collected on consumer attitudes and behaviour, via surveys. Despite the fact that there is a wealth of marketing data available from the separate surveys, reports are generally only created for each source individually. No single source of comprehensive information is available for in-depth data mining that can assist in identifying business opportunities (Van der Putten, Kok and Gupta, 2002). As a result marketers often request more detail in their consumer surveys to address all their research needs in a single source.

This need for information in survey research places a large demand on the consumer to provide accurate and detailed information on attitudes and behaviour through the use of longer questionnaires. Consequently the quality of responses is affected through respondent fatigue and even an increase in survey non-response due to refusal to participate in time consuming surveys (Raghunathan and Grizzle, 1995).

One possible solution to this problem of questionnaire overload is to divide the larger survey into smaller parts and administer each part to different samples from the same target population (Raghunathan and Grizzle, 1995). The separate databases would then be combined through data integration.

The Advertising Research Foundation (ARF) defines data integration as follows:

> *"A formal process to combine information from two or more separate data sources, making use of information of the databases for the purpose of accurately estimating values that are not available in any single data source".* (ARF, 2003)

Data fusion, also called statistical matching or synthetical matching, is a data integration technique used for linking multiple data sources through a set of common characteristics (D'Orazio, Di Zio and Scanu, 2006). The information in the individual data sources is collected from different but similar respondents from the same target population.

Consider the situation where two independent surveys are conducted on similar respondents within the same target population, say survey A and survey B. Both surveys include a set of common characteristics **X** that are comparable between the two surveys. Survey A further includes a set of measurements **Y** while survey B consists of a different set **Z**. As a result the information for **Z** is not observed in survey A and **Y** is not observed in survey B. **Y** and **Z** are referred to as the variables unique to each source. The pattern of observed and unobserved data for this scenario can be illustrated as follows, where the shaded areas represent observed values and the blank areas the missing or unobserved data:

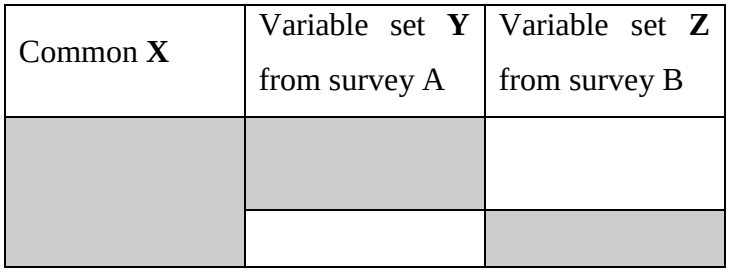

Figure 1.1: Illustration of data fusion

This illustration clearly shows that no information regarding the joint distribution of variables **Y** and **Z** is available, since these were never jointly observed. The objective of data fusion is to estimate the joint distribution of **Y** and **Z** using the information collected in the two independent samples. This will enable the analyst to construct a synthetic data file, linking sets of information that were never jointly observed. It therefore creates a "complete" data file that contains all the information from the separate data sources, as if the entire survey was administered to each respondent.

Data fusion can only be used as a viable solution to the problem of questionnaire overload if it will result in a valid data set that reflects the true relationships between the variables of interest. This largely depends on the link between the set of common variables and the unique variables, i.e. the underlying mathematical model that defines the bridge between the individual data sources. However, since the unique variables are never jointly observed, this link requires certain assumptions that are generally impossible to test in practice. The most common model used to fuse data is based on the assumption of conditional independence (CIA), where **Y** and **Z** are assumed to be independent, given knowledge of **X**.

Any data fusion application must always be evaluated for its accuracy and validity to determine if the fusion was successful. For data fusion done on real data, this exercise is restricted to the information that is already available in the data sources used. Rässler (2002) defines four levels of validity, namely evaluation of the preservation of individual values, joint distributions, correlation structures and marginal distributions.

1.2 Problem Statement

Much of the data fusion research involves the use of real data sets. The variables in the data could be continuous, categorical, or both, depending on the industry and the nature of the research. Examples of this can be found in O'Brien (1991), Tchaoussoglou and Van der Noort (1999) and Soong and De Montigny (2001).

The fusion process is frequently evaluated through simulation (D'Orazio *et al*, 2006), where data are simulated from a specific distribution with pre-specified parameters that represent real world situations. Such data are seen as the theoretical complete data set and are randomly divided into subsets, which are then fused together through a particular mathematical process and under certain assumptions. To assess the success of the fusion, the distribution of the variables in the fused file is compared with the distribution in the original simulated data set.

In most simulation studies the focus has been on generating continuous data, such as multivariate normal random variables, in order to study distributional properties. This is done by either creating only one synthetic data set (single imputation) or multiple fusions for the same data (multiple imputation). For the latter, the desired analysis is performed on all fused data sets and the output is combined into a single estimate for the analysis. For example, Rässler and Fleischer (1998) perform various single imputation data fusions on simulated normal and lognormal data, while Kiesl and Rässler (2006) use simulated data from the multivariate normal distribution to assess a multiple imputation algorithm for data fusion.

Market research data are typically collected via consumer surveys and the variables are generally categorical, such as Likert-scale attitudinal measurements, or binary responses to media consumption and product usage. For example, consider a survey consisting of a number of categorical variables ($\mathbf{X, Y, Z_1, Z_2}$). Variables $\mathbf{X}$ record demographic, geographic and key behavioural information of the target population. Media consumption is measured through a set of indicator variables $\mathbf{Y}$, while $\mathbf{Z_1}$ and $\mathbf{Z_2}$ denote product usage for two different product categories.

In practical data fusion applications this survey would be subdivided into two separate surveys and administered to different respondents from the same target population, where surveys A and B consist of variables ($\mathbf{X, Y}$) and ($\mathbf{X, Z_1, Z_2}$) respectively.

Simulation of artificial data that represents such a survey consisting of categorical variables will provide the analyst with a useful tool to investigate the feasibility of data fusion in the market research context.

The objective of this study is to evaluate the process of data fusion for binary data under the assumption of conditional independence, and to assess how deviations from

this assumption will impact the results. The research questions that will be addressed in this report are:

1. Can the level of conditional independence in binary data be quantified?
2. How successful is binary data fusion for different levels of conditional independence in the data?
3. Is the success of a binary data fusion dependent on the strength of the relationship between the unique variables?
4. To what extent is the fusion successful if conditional independence is valid for only a subset of the unique variables?

In order to address these questions, data will be simulated to reflect the distribution of survey-based data, where nominal and ordinal variables are represented as binary indicator variables. The simulation is based on a pre-specified marginal distribution and correlation structure, using the binary simulation technique proposed by Alosh and Lee (2001). Since results from a single fusion exercise may appear "good" purely by chance, a single simulation will not be sufficient to address the questions posed above. To avoid this problem, a total of 30,000 binary data sets will be simulated and the fusion performed on each set.

Any simulation that involves a large number of variables adds to the computational complexity of the analysis. For this analysis, the survey described above will be reduced to only four binary variables (X, Y, Z_1, Z_2). This set will be sufficient to assess the success of binary data fusion under CIA.

The data will also be simulated to reflect varying levels of conditional independence in the data, ranging from complete conditional independence to the absence thereof. The degree of conditional independence in the data will be quantified using a function of entropy, called the conditional mutual information (CMI) of a distribution. In addition to this, the effect of the strength of the relationship between the unique

variables on the quality of the fusion will also be evaluated. Thus, the simulated data will also display weak to strong levels of correlation between the unique variables.

Each simulated data set will be randomly divided into two subsets of approximately equal size and the fusion algorithm applied for each simulation, as illustrated in Figure 1.1. The fusion algorithm estimates the complete joint distribution of four binary variables using the information in the two random subsets. It is based on maximum likelihood estimates (MLE).

The success, or failure, of each fusion will be evaluated through a series of statistical tests, comparing the fused data set with the original simulated data set. This evaluation is within the framework proposed by Rässler (2002) to assess the validity of a fusion. The results from the various tests will provide an indication of how deviations from conditional independence impact the success of binary data fusion.

The R software package is used for all simulations and statistical tests. The authors of the software state that "*R has a home page at http://www.R-project.org/. It is free software distributed under the GNU-style copyleft, and an official part of the GNU project (GNU S)*".

1.3 Structure of the Report

The report is structured as follows: Chapter 2 provides a detailed theoretical outline of data fusion, specifically regarding modelling approaches, imputation methods, linking algorithms and model assumptions. Chapter 3 reviews the literature concerning missing data, applications of data fusion, and binary data simulation. Chapter 4 outlines the methodology for the binary simulation, the data fusion, and evaluation of the quality of the fusion. The analysis is presented in Chapter 5, with conclusions and recommendations in Chapter 6.

2 THEORETICAL OUTLINE

2.1 Concepts and Definitions

In practical applications there are two distinct forms of data integration: record linkage, also called exact matching, and statistical matching or data fusion (Radner, Allen, Gonzalez, Jabine, and Muller, 1980). It is important to clearly distinguish between these two forms, as the data structures and methods involved in the integration processes are quite different.

In record linkage the units of the different data sources are at least partially overlapping and are identified through a unique key such as ID number, tax number, name and surname, etc. (D'Orazio *et al*, 2006). The techniques used to integrate such data are focused on the appropriate identification of distinct units. If the quality of the separate data sources is such that individuals are clearly identified, a very accurate database can be constructed from these data files.

A key concern with record linkage as a data integration technique is the issue of confidentiality. This matter was already a barrier during the early stages of exact matching applications. Radner *et al* (1980) state that USA legislation, such as the 1974 Privacy Act and the 1976 Tax Reform Act, further restricted the use of record linkage as a feasible solution for data integration, and that many statistical agencies have statutes that prohibit them from disclosing personal information collected on individuals in surveys.

More recent record linkage techniques have also been used for purposes other than its original application. Herzog, Scheuren and Winkler (2010) describe how advances in computer technology lead to the development of record linkage models that can aid in improving the quality of business or government sampling frames by identifying

duplicate records. By removing such duplicates the sampling frame better represents the population of interest. Furthermore, it eliminates the possibility of selecting the same response unit more than once.

The restrictions on the use of record linkage for combining data sets sparked an increased interest in further development of alternative methods of data integration, particularly that of data fusion.

Although data fusion also involves linking data from separate sources, the most notable difference between data fusion and record linkage is that the units of the different data sources do not overlap, or at least that it is not possible to uniquely identify corresponding respondents in the input data sets. In general, the units (or respondents) in this case are regarded as different but similar individuals from the same target population. Records are linked, not through a unique identifier, but through a set of common characteristics that identify those respondents that are most similar (D'Orazio *et al*, 2006). The common characteristics could be continuous or categorical.

The common characteristics, also called the matching variables, are classified as either critical or non-critical variables (Soong and De Montigny, 2001). Critical variables are defined as all variables for which an exact match is required, such as gender. This means that data for males will only be fused with other male respondents in the sample, similarly for female respondents. Nominal variables that form part of the common set are generally used as critical variables.

Non-critical variables are typically ordinal or numerical variables for which an exact match is not essential or even possible. For example, a respondent aged 30 can be fused with another respondent of similar age, say 32 years old. These two respondents are seen to be similar enough in terms of their ages, so that they can be linked together without any loss of information. Other examples of non-critical

matching variables include income, number of children in the household and product ratings on an ordinal scale.

The data from the source files are first divided into subsets defined by the levels of the critical variables, such as the two gender groups. Within these levels the data are typically linked using a distance measure, such as the Euclidian distance between the two vectors of non-critical matching variables. Records are then fused on the closest match (Rässler, 2002). D'Orazio *et al* (2006) also describe methods of fusing data using parametric models such as the regression of the variables unique to each of the two data sets on the common variables, to determine the link between the variables not jointly observed.

In practical applications, data fusion is carried out for two different research situations, namely designed fusion and ad-hoc fusion (Galpin and Neethling, 2004). In designed or planned fusion, a survey questionnaire is divided into smaller parts before it is administered to the different groups. This is the most desirable approach as the researcher decides upfront how the questionnaire must be divided and which variables will be used as the common variables. This will also ensure that the set of common variables is formulated in the same way. The separate data sources are therefore aligned in terms of the variables as well as the population units.

Ad-hoc fusion is a method of integrating large amounts of data that were collected independently. In such fusion projects, the researcher is restricted to the set of variables that are found to be common to all the individual data sources. To ensure that the files are comparable, it may be necessary to make some adjustments to the variables, such as combining age or income categories so that both sets use the same definition. For individual surveys, sample weights are used to adjust the sample distribution of the variables used to design the sample to match the population distribution. If the population distribution for independent surveys used in the fusion differs, the sample weights should also be rescaled to a common population total.

Despite the considerable amount of data preparation necessary for ad-hoc data fusion, it is still a viable solution that can potentially produce a more holistic view of the research objective.

2.2 Data Fusion as a Missing Data Problem

At the heart of the data fusion problem is the occurrence of missing data in sets of information, as illustrated in Figure 1.1. In order to understand the data fusion process it is important to examine the causes of missing data, the impact of this on the analysis, and generally how the problem of missing data is addressed in survey research. This will guide the use of appropriate techniques for specific situations to ensure that missing data is dealt with in the correct way.

2.2.1 Non-response

Missing data is also referred to as non-response. According to Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau (2009), non-response in survey research is described as "*the failure to obtain measurements on sample units*".

Two types of non-response can occur in survey data, namely item non-response and unit non-response. Unit non-response occurs when a unit drawn from the sample is not contactable and is thus not interviewed, therefore none of the survey information is measured on the selected respondent. Non-response can also appear in individual survey questions. This is termed item non-response or partial failure to obtain measurements (Groves *et al*, 2009).

Unit and item non-response are prompted by different events. Groves *et al* (2009) list the main causes of each type of non-response. For unit non-response, these include the event that a sampled person cannot be contacted, or the sampled person is unable to provide information, often due to a language barrier, or the respondent simply declines to participate in the research. Item non-response typically occurs when the

respondent does not fully understand the question, or feels unable to provide accurate answers to the question. For sensitive questions, such as a person's income, the respondent may be unwilling to disclose the information, also resulting in item non-response.

Groves *et al* (2009) state that non-response in survey data could negatively impact the quality of statistical analysis performed on the data. This is referred to as non-response bias and is due to the fact that the information gathered from respondents could differ systematically from the information that the non-respondents would have provided, if they were measured. For both unit and item non-response, the impact of non-response bias on survey estimates is effectively the same, though in item non-response the bias only affects the items in question.

Unit non-response is generally managed through the use of sampling weights, while item non-response is dealt with using either single or multiple imputation. For the latter, there are many different opinions in the literature as to which method leads to the best result. Regardless of the method used, it is important to thoroughly investigate the underlying missing data mechanism of each variable that is subject to item non-response in order to understand the causes of the missingness.

2.2.2 Missing data mechanisms

Little and Rubin (1987) formally define the notation and terminology of the three mechanisms that cause data to be missing. This is based on the relationship between the missing data and the values of observed data.

Consider a data set with variables X and Y. Assume that X consists of complete records whereas Y is subject to non-response. If the probability that Y is missing is unrelated to the values of both X and Y, the missing data is said to be Missing Completely At Random (MCAR). Several factors can result in this mechanism, such as equipment malfunction, incorrectly entered data, or the respondent was not asked

the question. The latter can occur by mistake, for example missing a page in the questionnaire, or by design, as in data fusion.

Suppose a certain subset of respondents is less likely to answer a particular question Y, but that the probability of non-response within that subset is unrelated to the value of Y, then the missing data mechanism is Missing At Random (MAR). For instance, suppose that variable X indicates whether the respondent is the primary household caregiver or not, and variable Y records the amount of laundry detergent used per month. Household members that are not involved in general household duties may be less likely to respond to the question about laundry detergent, therefore the missingness in variable Y is a function of the response to another variable X.

If the missing data mechanism is neither MCAR nor MAR, then is it said to be Missing Not At Random (MNAR). In this case the probability that Y is missing is related to the value of Y itself. A typical example of this is where variable Y records a person's income. High income earners are inclined to refuse to answer questions related to income, whereas people with low or average income are generally more willing to provide such information. In this situation the missingness in Y is due to the values of the variable itself.

For the data fusion problem, separate samples taken from the same population are linked together to form a complete data file. It is assumed that all the information in the separate sources is generated from the same joint population distribution, although sections of the data are not recorded due to unasked questions. The mechanism that caused the data to be missing is independent of both the observed and the missing data since the missingness is induced by design. It therefore follows that the missing data mechanism in the context of data fusion is MCAR (D'Orazio *et al*, 2006).

Due to its design, data fusion is only concerned with unit non-response. It is however possible that item non-response can occur in the individual data sources. This must be dealt with during the data preparation phase, prior to the fusion, to ensure that the information in the separate data files is as complete and unbiased as possible.

2.2.3 Dealing with missing data

A thorough understanding of the mechanisms that lead to missing data will enable the analyst to deal with the non-response in a way that should produce unbiased parameter estimates. A number of different techniques are used in practice and usually form part of standard statistical software.

The traditional techniques used to impute missing values include the deletion of records (listwise or pairwise), mean substitution, regression substitution and various hot-deck imputation methods (Howell, 2009; Lohr, 1999). Listwise deletion involves the deletion of an entire case. Under the pairwise deletion approach, all available data are used in the calculation of the correlation matrix. If a respondent has a missing value on only one variable, the rest of the data for that respondent are still used in calculating the correlations between the observed variables.

Howell (2009) defines mean substitution as replacing all the missing data for a particular variable with the mean value of that variable, while regression substitution involves predicting the values of the missing data for a specific variable through the linear relationship between that variable and a set of other related variables. Hot-deck imputation methods use information from similar respondents, identified through nearest-neighbour algorithms, to impute missing data.

More modern approaches are maximum likelihood procedures such as the Expectation-Maximization algorithm (EM) and multiple imputation. The EM algorithm is an iterative method for finding the MLE in a two-step process (Rässler, 2002). Missing data are first estimated by their conditional expectation, given the

observed data, and the initial MLE for the parameters of interest. In the second step the MLE are updated using observed data as well as the imputed data. Both steps are repeated until the estimates converge.

Rässler (2002) defines multiple imputation as a technique whereby each missing observation is replaced with several plausible values, say *m*, resulting in *m* complete data sets. The imputed values are randomly selected from a distribution that reflects some assumptions about the missing data in an attempt to represent the true distribution of the variable of interest. The desired statistical analysis is then performed on all *m* data sets and the results combined to provide one set of parameter estimates and confidence intervals for the analysis.

2.3 Data Fusion Considerations

Before an analyst can perform data fusion on a set of data files, a number of different aspects must be considered. This involves decisions on the appropriate modelling approach, imputation methods, linking algorithms and model assumptions.

2.3.1 Macro vs. micro modelling

According to D'Orazio *et al* (2006), data fusion can be approached in two ways. If the focus is on the joint distribution of **Y** and **Z**, or special characteristics thereof such as the correlation matrix, macro modelling would be sufficient to estimate the required parameters without the need to link individual records. However, most practical applications of data fusion aim to create a synthetic data file at respondent level. This is called micro modelling and provides the analyst with much more flexibility to perform a wide range of statistical analyses.

2.3.2 Parametric vs. non-parametric modelling

Certain assumptions about the joint density function $f(x, y, z)$[1] of the random variable (**X**, **Y**, **Z**) determine the modelling framework for data fusion (D'Orazio *et al*, 2006). If $F = \{ f \}$ is assumed to be a parametric family of distributions then the density $f(x, y, z)$ is defined by a finite set of parameters Θ. In this framework parametric methods for estimating parameters can be used to gain information about the joint distribution of (**Y**, **Z**) for both macro and micro approaches. If no assumptions are made about the distribution of (**X**, **Y**, **Z**) then the modelling is performed using techniques defined in a non-parametric framework, such as hot-deck imputation methods.

2.3.3 Single vs. multiple imputation

Since data fusion is a missing data problem, it involves the imputation of the missing observations through the modelling of observed data. This can be done through either single or multiple imputation. In single imputation, each missing observation is replaced with a single value, while multiple imputation involves replacing each missing observation by two or more acceptable values, randomly drawn from a distribution that represents a set of possible values for the missing observations (Rubin, 1987; Rässler, 2002).

Because single imputation does not account for the uncertainty regarding the parameters estimated from statistical analyses using filled-in data, multiple imputation addresses this problem and provides a way to represent the uncertainty associated with imputed values. Multiple imputation was first introduced by Rubin (1977) to handle item non-response and was generalized as a data fusion procedure by Rässler (2002).

[1] For ease of notation $f(x, y, z)$ refers to the probability density function for continuous variables and the probability mass function for discrete variables

2.3.4 Constrained vs. unconstrained fusion

In the context of data fusion two types are defined, namely constrained and unconstrained methods (Radner *et al*, 1980). In unconstrained matching one of the source files is seen as the donor and the other as the recipient. Records from the donor file are matched with one or more records of the recipient file, therefore the number of records in the fused file is the same as the number of records in the recipient file. As a result the sample weights and original distributions of the recipient file are preserved. The records from the donor file could be used multiple times without constraint or weight adjustment. Consequently the data in the donor file are re-weighted, which may lead to some distortions in the original distributions.

The algorithm for linking one donor to one recipient in an unconstrained match, searches through the donor data file for the single record that is similar to the recipient record. This process is repeated until each recipient record has a match. The major disadvantage of this approach is that it could fuse records that are very dissimilar, purely because the closest record has already been assigned to a recipient. This can be controlled by specifying the maximum distance allowed for successful matches. When fusing a single donor record with multiple recipient records, a penalty function must be incorporated to ensure that records are not used too many times in the fusion (Rässler, 2002).

In constrained matching all the records from both data sources are used to create the synthetic file in such a way that all the original marginal and joint distributions as well as the weights are preserved (Radner *et al*, 1980). This can be achieved by minimizing the weighted distance between all records in the individual data sources, subject to constraints imposed on the weights (Rodgers, 1984). The transportation algorithm is often used in constrained data fusion applications.

The transportation algorithm is a linear programming model that is used as an optimization algorithm in operations research (Sivazlian and Stanfel, 1975). The

main idea is to minimize the cost of transporting goods from a number of sources to a number of destinations. Each source has a known supply of the goods available and the destinations have a limit as to the quantity of goods that can be stored on site. The cost of transportation between each source and each destination is also known.

This technique is used in data fusion to implement an optimal constrained match between two data sets taken from the same target population. These files are weighted to the same population totals and the weights can be viewed as the supply and demand that must be transported and assigned in the fusion. The distance between records for the set of common characteristics can be seen as the cost of the transportation. This approach will ensure that all records from both files will be used in such a way that the original weighted marginal and joint distributions are preserved.

2.3.5 The problem of identification

The identification problem is inherent to data fusion and concerns the joint distribution of variables **Y** and **Z**, which are not jointly observed. The objective of the fusion is to estimate the true relationship between **Y** and **Z** through the marginal and joint distributions in the individual data sources. However, these distributions alone do not provide sufficient information to uniquely identify the joint distribution that could have generated the data and additional assumptions about the data must be made to ensure that the joint distribution can be identified (Gilula, McCulloch and Rossi, 2006).

According to D'Orazio *et al* (2006), only a small number of identifiable models exist that can accurately estimate the parameters of the joint distribution (**X**, **Y**, **Z**). These models require certain assumptions about the underlying distributions of the data, and include the conditional independence model, the pairwise independence model and finite mixture models.

Alternative approaches that are used to overcome the problem of finding, as well as evaluating identifiable models, include models such as the Bayesian regression approach using multiple imputation, as described by Rässler (2002). Auxiliary information can also be used to evaluate the assumptions of conditional independence and improve the model (Paass, 1986).

2.4 Identifiable Models

The conditional independence model is frequently used as the identifiable model in single imputation data fusion. This is evident from the vast number of applications found in the literature. However, the primary criticism about data fusion is directed at the feasibility of using the restrictive assumption of conditional independence (Rodgers, 1984). Rässler (2002) and D'Orazio *et al* (2006) provide comprehensive information on the background and mathematical derivations for the conditional independence assumption model.

Although this is not the only identifiable model for data fusion, it is the easiest model for estimating the joint distribution of **Y** and **Z**, provided that the assumption is valid. Alternative models are described in D'Orazio *et al* (2006) and the mathematical basis, assumptions and estimation procedures for each identifiable model are discussed in the following sections.

2.4.1 Conditional Independence Assumption (CIA)

The CIA is an identifiable model used for fusing data through single imputation and makes the explicit assumption that **Y** and **Z** are independent, given the set of common variables **X**. This implies that all the information about the relationship between the unique variables **Y** and **Z** is transmitted through the information contained in the set of common variables **X**.

Using properties of conditioning, the joint density function (or the joint probability function) can be written as follows

$$f(x, y, z) = f_{Y,Z|X}(y, z \mid x) f_X(x). \tag{2.1}$$

Under the assumption that **Y** and **Z** are independent given **X**, the joint density in equation (2.1) can be further simplified and expressed in terms of the conditional densities of **Y** given **X**, and **Z** given **X**. Therefore, the joint density can be written as

$$f(x, y, z) = f_{Y|X}(y \mid x) f_{Z|X}(z \mid x) f_X(x). \tag{2.2}$$

If the assumption of conditional independence holds true, the joint density can then be completely identified through the conditional and marginal distributions from the separate data sources, through equation (2.2).

2.4.2 Pairwise Independence Assumption (PIA)

Under the CIA model, **Y** and **Z** are assumed to be independent, given knowledge of **X**. It is however possible that **Y** and **Z** are marginally independent, but that the introduction of **X** creates dependence between **Y** and **Z**. This dependence structure leads to the pairwise independence assumption (PIA), another identifiable model for single imputation data fusion.

Consider the 3-variate categorical distribution (X, Y, Z) with categories (i, j, k), where $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, such that $\theta_{ijk} = P(X = i, Y = j, Z = k)$. Furthermore, assume that Y and Z are marginally independent, but Y and Z given X are dependent. This dependence structure can be defined by a log-linear model with a zero three-way interaction term, namely

$$\log\left(n\theta_{ijk}\right) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{2.3}$$

The objective is to find the MLE for $\hat{\theta}_{ijk}$ in order to determine the complete joint distribution of (X, Y, Z) from the two separate data sources, where file A consists of X and Y, and file B consists of X and Z. The minimal sufficient statistics to estimate the model given by equation (2.3) are the marginal tables for (X, Y), (X, Z) and (Y, Z). Because Y and Z are assumed to be marginally independent, the marginal tables for (Y, Z) can be estimated from the individual data sources. Under the PIA model, $\hat{\theta}_{ijk}$ cannot be found in closed form and the iterative proportional fitting algorithm is used to estimate $\hat{\theta}_{ijk}$.

2.4.3 Finite mixture models

The finite mixture model is a form of the CIA model that is based on the assumption that **X**, **Y** and **Z** are independent, given a latent variable ***l***. The latent variable is missing in both data sets A and B, but if it can be determined, the association between the variables are assumed to be independent. The latent variable consists of a mixture of G different distributions and the mixture depends on the mixing proportions π_l such that $\sum_{l=1}^{G} \pi_l = 1$ and $\pi_l \geq 0$. Therefore, the joint distribution function can be defined as

$$f(x, y, z; \theta) = \sum_{l=1}^{G} \pi_l f_X(x; \theta_l) f_Y(y; \theta_l) f_Z(z; \theta_l). \tag{2.4}$$

Estimates for the parameters $(\hat{\theta}, \hat{\pi})$ in equation (2.4) are generally found through the EM algorithm.

2.4.4 Auxiliary information

One possible solution to the problem of assumed conditional independence is to use auxiliary information from another data source that contains partially observed information about the true relationship between **Y** and **Z**. This concept was first

suggested by Paass (1986) and involved the use of additional constraints based on the external source to improve the quality of the fusion. Such information is often in the form of frequency tables, regression equations or covariance matrices.

It is important that the auxiliary information is sourced from the same population as the data sources to be fused. This may not always be the case, for example, samples taken at a different point in time, or from different geographic regions. Paass (1986) stressed that auxiliary information should only be used if there is sufficient justification that it would improve the fusion results.

2.5 Data Fusion Methods under the Assumption of Conditional Independence

D'Orazio *et al* (2006) describe a set of different approaches for data fusion under the assumption of conditional independence, specifically with respect to the modelling framework. The following sections provide details for each of these models for the trivariate distribution (X, Y, Z), as outlined in D'Orazio *et al* (2006). Two separate data files (A and B) are used to estimate the overall distribution. These models can be extended to multivariate distributions.

2.5.1 Macro parametric modelling

Consider the situation where $F = \{ f \}$ is assumed to be a parametric family of distributions. Therefore the density $f(x, y, z)$ is defined by a finite set of parameters Θ. Under the assumption of conditional independence, the joint density can be expressed as follows

$$f(x, y, z; \theta) = f_{Y|X}(y \mid x; \theta_{Y|X}) f_{Z|X}(z \mid x; \theta_{Z|X}) f_X(x; \theta_X).$$

For the macro approach to data fusion it is sufficient to estimate the parameters θ_X, $\theta_{Y|X}$ and $\theta_{Z|X}$. Since a parametric model is assumed, these parameters can be

estimated through maximum likelihood procedures. The MLEs for θ_X, $\theta_{Y|X}$ and $\theta_{Z|X}$ are computed from the combined sample, subset A and subset B, respectively.

2.5.2 Micro parametric modelling

If a parametric model can be assumed (and estimated) for the data, then the missing observations for all respondents can be predicted from the plausible values defined by the specific parametric distribution. D'Orazio *et al* (2006) describe the two main fusion methods in a micro parametric framework: conditional mean matching, and draws based on a conditional predictive distribution.

Conditional mean matching

The most commonly used method to impute missing data in the micro parametric data fusion context is to replace each missing value of a variable with the expected value of the variable. If the continuous variables X, Y and Z are assumed to be normally distributed, then the missing information for Z in file A, and the missing information for Y in file B can be replaced with the estimated regressions of Z on X and Y on X respectively. A disadvantage of this approach is that the predicted value may not be valid in terms of the scale of the variable. Furthermore, the resulting fused file will have data for variables Y and Z that are concentrated around the conditional mean, which will influence the variance of the fused distributions.

Draws based on a conditional predictive distribution

This approach is an improvement on the conditional mean matching approach. Under the assumption of normality for variables X, Y and Z, the regressions of Y on X and Z on X are also used in this model. As before, the missing information in each file is predicted through the estimated regression models, but a random error term is added to the regression equations in order to better account for the variability in the original data files A and B. The error terms for imputing missing data in file A and B follow

a normal distribution with zero mean and estimated residual variance $\hat{\sigma}^2_{Z|X}$ and $\hat{\sigma}^2_{Y|X}$ respectively.

2.5.3 Macro non-parametric modelling

D'Orazio *et al* (2006) state that the multinomial distribution is a very flexible parametric model to use when fusing categorical variables. It is more difficult to define and to estimate the underlying parametric distribution for a set of continuous variables, especially in the multivariate context. Non-parametric modelling is therefore an appealing alternative to parametric modelling approaches, since it does not restrict the analysis to any distributional assumptions in the data.

As with macro parametric modelling, the objective of its non-parametric counterpart is to estimate the joint distribution (X, Y, Z) under the assumption of conditional independence. The difference is that non-parametric procedures are used to estimate the parameters of the joint distribution given in equation (2.2). Such procedures include the kernel density estimator or the *k*-nearest neighbour (*k*NN) estimate of a density function. Non-parametric regression can also be used to model the conditional distributions of Y given X, and Z given X.

2.5.4 Micro non-parametric modelling

In contrast to macro non-parametric models, micro level data fusion under the non-parametric framework has been extensively used in practice. These models are generally referred to as hot-deck imputation procedures. D'Orazio *et al* (2006) describe the three different types that have been used in data fusion applications: random hot-deck, rank hot-deck and distance hot-deck.

Random hot deck

In some surveys, it may be that the set of common characteristics **X** consists of only categorical variables, such as gender, province, socio-economic status, marital status,

etc. If it can be assumed that the combination of the levels of these variables identify subgroups of respondents with similar behaviour, the set of **X** variables can be used to divide the individual data sources into homogeneous subsets, called donation classes. Within these classes, one or more donors are then randomly selected and assigned to a recipient.

Rank hot deck

When the common characteristics consist of ordinal variables, it is not possible to compute numerical distances based on the values of the ordinal scale. In order to find the closest match between respondents, the data in each file are ranked separately based on the values of **X**. When the sample sizes of the two data sources are the same, the files are fused by linking records with the same rank. If the sample sizes are different then the absolute rank value cannot be used to find the nearest neighbour. In such cases the empirical cumulative distribution function is used to identify the closest match, in other words, records with the nearest cumulative relative frequency value are linked together.

Distance hot deck

For continuous variables, numerical distance functions are used to identify the nearest neighbour. A number of different distance functions are available, namely the Manhattan metric (city-block), Euclidian metric, Mahalanobis metric and the Chebyshev metric. According to D'Orazio *et al* (2006) the Mahalanobis metric is the most popular distance hot deck method used in data fusion, though Ingram, O'Hare, Scheuren and Turek (2000) state that the Euclidian distance function is the most commonly used measure in defining distance between observations.

2.5.5 Mixed methods

D'Orazio *et al* (2006) state that in many data fusion applications, a combination of parametric and non-parametric procedures are used, rather than restricting the

analysis to only one approach. The mixed method approach therefore combines parametric and non-parametric methods in a two-step process. As a first step, the parameters of the joint distribution given in equation (2.2) are estimated in the parametric framework. Thereafter, a synthetic file is created using appropriate hot-deck methods, given the estimated parameters. An advantage of this second step is that the missing data are imputed using valid values of the scale.

For example, consider the trivariate continuous distribution (X, Y, Z). In the first step regression analysis can be used to estimate the parameters of the distribution under the assumption of conditional independence. Missing values are then predicted through the estimated regression equations. The second step of the process involves the use of nearest neighbour techniques to impute the missing data in a non-parametric framework. In particular, missing values are imputed with valid values that are closest to the predicted values from the initial parametric model.

2.6 Fusion Validity

The ARF "Guidelines for data integration" (ARF, 2003) emphasizes the need to provide some evidence of the quality of a data fusion application, before it can be viewed as successful. The validation process defined by Rässler (2002) was approved by the ARF in 2003 as a framework for testing the validity of a fusion. This process consists of four levels of validity:

Level 1: Preserving individual values

Level 2: Preserving joint distributions

Level 3: Preserving correlation structures

Level 4: Preserving marginal distributions

The conditions for achieving each of the four levels of validity are outlined below, following Rässler (2002).

2.6.1 Level 1: Preserving individual values

This is the most difficult level of validity to achieve, as the true but unknown values must be recreated through the fusion. In an unconstrained match the unknown **Z** values in the recipient file are imputed from the donor file. This will create a synthetic data file with the distribution $f(x, y, \tilde{z})$, where $\tilde{z}$ are the estimated fused values of the variable **Z**. In order to attain this level of validity, the true values $f(x_i, y_i, z_i)$ for each respondent i must be recreated from the imputed values $f(x_i, y_i, \tilde{z}_i)$. This is only possible if the common variables **X** completely determine the unique variables **Y** and **Z**.

2.6.2 Level 2: Preserving joint distributions

Since the fused data set is seen as a random sample drawn from the underlying distribution $f(x, y, z)$, the focus is more on estimating the true joint distribution of **X**, **Y** and **Z** than on preserving the individual values. For single imputation, this is only possible if the CIA is true. If this level of validity can be attained, valid statistical analyses can be performed on the fused data set.

2.6.3 Level 3: Preserving correlation structures

For macro modelling the objective is to ensure that specific characteristics of the true joint distribution (**Y**, **Z**) are preserved. The estimated covariance (or correlation) structure in the fused file is the same as the true covariance (or correlation) structure only if **Y** and **Z** are on average conditionally independent of the common variables **X**.

2.6.4 Level 4: Preserving marginal distributions

A minimum requirement for a successful fusion is that all marginal and joint distributions from the individual data sources must be preserved in the fused data set. In estimating the unknown joint information (**Y**, **Z**) it is important not to lose or distort any of the known information from the separate sources.

In practical situations it is only possible to test the fourth level of validity, as the other levels require knowledge of the true distribution $f(y,z)$. These levels are typically assessed with simulation studies, the split-sample method or the use of auxiliary data. In the split-sample method, data collected from a large survey are divided into two or more subsets, which are then fused together to create a synthetic data file (ARF, 2003). This is also referred to as the fold-over method. In general, a fusion is seen as successful if the fourth level of validity is attained (Rässler, 2002). Empirical distributions or moments are compared using χ^2-tests or t-tests and differences should not be larger than would be expected between moments taken from independent random samples from the same population.

2.7 Issues in Data Fusion

The major concern about the application of single imputation data fusion is that it relies heavily on the validity of the CIA (Rodgers, 1984). Through this model, the independence of **Y** and **Z**, given **X** is mathematically imposed on the data. However, this cannot be tested in an applied situation because the joint data do not exist. A fusion under CIA could be improved with the use of auxiliary data, as suggested by Paass (1986). Such external sources of information must be comparable with the original data sources in terms of the target population, measurement unit and time period of assessment. The auxiliary data should also be generated from the same distribution that generated the data in the individual sources.

The multiple imputation approach does not assume conditional independence but it does make assumptions about the multivariate distribution of the data. D'Orazio *et al* (2006) state that multiple imputation cannot be considered under the non-parametric framework and is therefore dependent on parametric distributional assumptions, such as multivariate normality.

The Bayesian approach to multiple imputation uses the density function of a pre-specified prior distribution in order to derive the posterior distribution (Rässler, 2002). For such models, normality is generally assumed in all the applications in the literature to ensure that an appropriate prior distribution can be identified. At a minimum, the data should be from a parametric family of distributions.

The choice of the common variables can determine whether the data fusion is successful or not (O'Brien, 1991). The mere fact that the variables are common to the individual data sources does not guarantee that these variables are relevant in establishing the desired link between the data and therefore will not produce a valid fusion.

The quality of the individual data sources also plays a very important role in the success of a fusion. Alter (1974) states that survey data are not error-free and that the fused data file can only be as good as the original data. He emphasizes the importance of a thorough quality assessment of any data used as input to a fusion, prior to the actual fusion application.

Despite all these concerns data fusion is still considered as an acceptable alternative to long questionnaires that negatively impact on the quality of the data, provided that any assumptions made are suitably justified.

3 LITERATURE SURVEY

The analysis involved in this report is based on a number of different concepts. The main focus is on data fusion in the market research environment. In order to understand the rationale behind data fusion it is important to evaluate the causes of missing data and the specific trends in survey research that contributed to the further development of data fusion techniques. Another core component of this analysis involves simulating binary data with a pre-defined structure. This chapter reviews the literature published regarding missing data trends, data fusion and binary data simulation techniques.

3.1 Causes of Missing Data

3.1.1 Literature review

In survey research, one of the most important objectives of the data collection phase is to persuade a randomly selected individual to partake in the research. Groves *et al* (2009) list the refusal to participate in interviews as one of the three key causes of unit non-response. This refusal may lead to biased results if people who are likely to refuse differ from people who agree to participate in the research. A sampled individual that refuses to participate has to be replaced with another individual. This leads to increased costs as it requires additional travel costs to reach the new sample individual and will also impact on the time taken to complete all interviews. The two remaining causes of non-response are non-contacts and a selected individual's inability to provide the desired information.

A number of factors can impact on a person's willingness to participate in a survey, such as questionnaire length, incentives, the effectiveness of the interviewer, as well as other social and security concerns on the part of the respondent. These causes and impact of non-response in survey research have been extensively studied. Groves,

Cialdini and Couper (1992) provide a detailed description of the various factors that influence a respondent's decision to participate in survey research. These included socio-demographic, survey design and psychological factors.

The socio-demographic characteristics of both the respondent and the interviewer can influence survey participation. It has been shown that the response rate can vary within different levels of a respondent's age, gender, income, health status, and so on. Groves *et al* (1992) claim that these factors are not causal to non-response, but that they create a psychological predisposition that could influence a person's decision. In addition to this, a respondent's past experience, together with some environmental factors such as crime rates can also impact on the decision to partake in the survey.

An interviewer's skills, experience, socio-demographic characteristics and interaction with the respondent also play an important role in convincing a respondent to participate. Another component stems from societal factors that shape perceptions about the value of survey-based research as well as the effect of constant surveying in everyday life situations. The design of the survey itself such as the length of the interview or the topic of discussion can also influence the decision.

Van der Noort and Tchaoussoglou (1995) examine the reasons behind the decline in readership levels, as found in the Dutch SummoScanner survey. They conclude that an increase in questionnaire length influenced the interviewers in such a way that they tried to rush through the questionnaire in an attempt to keep the interview as short as possible. This could potentially have a negative effect on the quality of responses.

De Heer (1999) compares response rate trends for Labour Force Survey data across sixteen countries, from 1983 to 1997. Although there were some large differences between the countries in terms of response rates and the various non-response components, the overall trend indicated a rise in the number of refusals over time. De Heer states that this was possibly as a result of societal changes and survey burden,

such as very long surveys or the opinion that there are generally too many surveys conducted in research. The impact of the survey topic on response rates was also investigated in this study using expenditure surveys from twelve countries. Overall, it was found that the response rate for such surveys was much lower than that of the Labour Force Surveys, which could be due to a higher response burden for expenditure surveys, since these are generally much longer and intensive.

Other research has shown that there is a general decline in willingness to participate in long questionnaires. Curtin, Presser and Singer (2005) used the data from the Survey of Consumer Attitudes (SCA) from the University of Michigan to assess the decline in response rate over a period of 25 years (1979 – 2003). They found that the response rate dropped from 72% to 48%, with the most notable change in recent years. From 1979 to 1996 there was an average decline in the response rate of 0.74 percentage points per year. After 1996 the response rate decreased by an average of 1.5 percentage points per year. They also found that the rise in survey non-response in later years was largely due to refusal to participate.

In an experiment to test the effect of interview length and incentives on respondents' willingness to participate, Hansen (2007) concludes that the length of a questionnaire had the biggest impact. When changing the announced interview time from twenty minutes to fifteen minutes, the number of completed interviews increased by 25%. Follow-up interviews with respondents who refused to participate showed that the excessive demand in survey research creates a reluctance to participate.

Groves *et al* (2009) state that non-response in government sponsored surveys is typically lower that non-response in academic surveys or those conducted in the private sector. Surveys administered by the US Census Bureau, such as the National Crime Victimization Survey (NCVS) and the Current Population Survey (CPS), both indicated a low but increasing non-response rate. Data from 1975 to 2007 for the NCVS showed a slight increase in both the overall non-response rate, as well as the

refusal rate, from about 1994. Despite the decline, the household response rate was still very high (96 – 97%).

They also state that the non-response rate in the CPS (1955 – 2005) remained consistent and generally low for many years, with only a small increase in the trend. However, in 1994, there was a noticeable change in both the non-response rate and the refusal rate. A possible explanation for this was the fact that the survey methodology changed from paper questionnaires to computer assisted interviews. The authors suggested that, since it was perhaps no longer possible to perform an interview "at the doorstep" using a computer, respondents were more inclined to refuse to participate in the survey.

3.1.2 Summary

According to Singer (2006), research into survey non-response over time went through various stages: from the initial investigations into whether response rates were actually declining or not, how widespread the decline was if indeed it existed, and which components of non-response were mainly affected. A number of authors found that non-response appeared to be rising in recent years and that this was often a function of respondents' refusal to participate in survey research.

Based on such findings, a great deal of research was undertaken into developing fieldwork procedures aimed at reducing all components of non-response (Singer, 2006). This included issues such as ethics, how to deal with an answering machine, the use of incentives and a considerable focus on interviewer training. The use of multiple imputation techniques to adjust for both unit and item non-response also received renewed interest. Singer (2006) further claims that the current atmosphere around this topic reflects a reluctant realization that the non-response phenomenon is escalating, and this trend is likely to continue in the future, despite all efforts to minimize non-response rates. The overall message from all the research is clear: survey response rates have declined over the years.

3.2 Data Fusion

3.2.1 Literature review

Radner *et al* (1980) note that the first application of data fusion was in the field of economics, done by the Bureau of Economic Analysis (BEA) of the U.S. Department of Commerce in 1968. The 1964 Internal Revenue Service (IRS) Tax Model (TM) of filed tax returns was fused to the 1965 Income supplement of the CPS. The purpose of this fusion was to improve the accuracy in estimating the size distribution of family personal income. Some universe and unit adjustments were made to both sources to ensure alignment.

Records from the TM were linked with CPS records in a constrained match based on the size of income within cells defined by the critical variables, by ranking the income levels in each of the two surveys. Weights for both files were preserved by duplicating records and splitting the weight. The set of common variables, both critical and matching, were subjectively chosen based on their presumed explanatory power. A Pilot Link Study in 1963, that merged information through record linkage, was used to derive the order of importance of the common variables.

In 1969 this fusion was further enhanced with additional information about specific income types from the 1962 Survey of Financial Characteristics of Consumers (SFCC). The SFCC contained income, asset and liability data for approximately 2500 households. As with the 1965 CPS-TM file, the fusion was based on ranking of size of income within cells identified from variables that were assumed to be related to the additional SFCC income types. For this study the unconstrained approach was used and SFCC records were re-weighted to align with the weights within cells in the CPS-TM.

Another first generation data fusion application, described by Okner (1972), was performed by the Brookings Institute. This fusion is known as the MERGE-66 and

involved a match between the Survey of Economic Opportunity (SEO), conducted in 1967, and the IRS Tax File (TF) of individual federal income tax returns of 1966. The SEO consisted of demographic data and claimed income information for 1966 for a sample of approximately 30,000 households. The objective of this project was to create a more detailed micro-level data file that could be used for tax policy analysis. The TF was used as the donor and the SEO the recipient in an unconstrained match.

The first step in the MERGE-66 fusion was to create groups in both files based on income type, marital status, age and number of exemptions for dependents. A consistency score was then calculated within each group where points were allocated to variables that indicate major and minor income sources. Only records with the highest 25% of consistency scores within groups were linked together based on the closest match. For difficult matches the criteria for defining a "nearest neighbour" was somewhat relaxed in an attempt to link as many records as possible. However, 97% of the TF records were fused to the SEO file using strict criteria. Some universe adjustments were made in both files to align the populations.

After the MERGE-66 file was created, the authors were confronted with some distributional issues. Substantial differences were found between the derived income distribution and that which was published by the IRS for high-income earners. This was due to the SEO sampling scheme, where very few high-income families were sampled, reflecting the general population. Contrary to this, the TF included a large number of tax returns filed by the high-income earners. The analysts attempted to solve this problem by splitting the MERGE-66 file into two subsets: all families with an income less than $30,000, vs. high income families. The fused file for the first group was retained, but the income information for the remaining group was replaced with the TF data. This implied that there would be no SEO demographic data for the second group.

Also, the income in the fused file was less than the 1966 Adjusted Family Income (AFI), derived from personal income figures of the Office of Business Economics and IRS data. The analysts therefore had to adjust the MERGE-66 income information to reflect the national AFI figures, by applying a ratio to the reported income in the fused file. Overall the authors felt that, despite all the difficulties and distributional issues in this fusion exercise, the MERGE-66 file was very useful for analytical purposes.

Budd (1972) comments on the MERGE-66 data fusion procedure. His main criticism refers to the issue that claimed income in survey data is typically under-reported. As a result, tax returns in a certain income class were fused to SEO records that were probably in a higher income class. Although some adjustments were made to the income distributions in the MERGE-66 file, this was done after the fusion, and was not sufficient to deal with bias due to under-reporting.

According to Radner *et al* (1980), most of the initial data fusion applications were done with existing data sources and population or unit adjustments were often required to align the data. No, or very few, additional sources were available to evaluate the quality of the fused files. The choice of the common variables X and their relative importance was often subjective and not based on any statistical analyses. Scores were calculated based on the importance of the common variables. Non-parametric rank hot-deck algorithms were mainly used to link records in both constrained and unconstrained approaches. For constrained fusion, sample weights were typically split, based on rankings within critical variables.

Sims (1972) addresses the underlying mathematical models for data fusion and observed that the joint distribution of X, Y and Z after the fusion is only equal to the true joint distribution if the partial correlations among the sets of Y and Z variables given X is zero. He states that the main assumption for a fusion to be valid is that of conditional independence.

Alter (1974) describes a planned fusion application of the 1970 Canadian Survey of Consumer Finances (SCF) and the 1970 Family Expenditure Survey (FEX). The main objective of this project was to measure and compare the relative income distribution of the Canadian population internationally using the United Nations System of National Accounts (SNA). Individually, the SCF and FEX did not contain all the information required to apply the SNA and using a single survey was not an option. It was felt that such a survey would cause extreme response burden, non-response and increased response error. The researchers therefore designed a fusion between the two surveys, where the FEX survey would be used as the donor and the SCF as the recipient.

The 1970 SCF was a national income survey with a special section on assets and debt, administered to approximately 10,000 families and unattached individuals. The 1970 FEX consisted of 14,000 families and unattached individuals, and covered detailed household consumption and expenditure. A core set of common questions were determined based on prior knowledge and in view of the objectives of the study. Both surveys were administered in early 1970 to the same target population, as defined by the Canadian Labour Force Survey. The whole design of the project ensured that the two surveys were as compatible as possible and that no sample weight or question alignments were necessary during the fusion.

The surveys were divided based on the two main critical variables, namely home ownership and family type (family vs. individual). Further critical variables included region, main income source, age category of the head of the family, gender and children in the family. The matching variables were determined within each subset through multiple linear regression analysis in an attempt to make the choice of the common variables more objective and to aid in deriving the relative importance of each variable in the set. In order to be considered, the matching variables had to show the same explanatory power in both asset holding and debt patterns (SCF) and consumption (FEX).

The final set of matching variables was rank-ordered according to importance, as derived through the regression analysis, and the relative impact was then quantified in terms of union scores. These were calculated as a measure on closeness between the records of SCF and FEX, based on the matching variables. The number of times that an FEX record could be used as a donor was restricted to 16. However, 43% of the records in the FEX file were used only once in the fusion.

The set of common variables from both data sets were compared in the fused file using cross-tabulation for categorical variables, and paired-sample t-tests for continuous variables. Overall the quality of this fusion was not completely satisfactory. Some of the results were promising while others were poor. Alter (1974) states that there is room for improvement in the fusion techniques used here. He noted that the quality of any data fusion also depends on the quality of the individual data sources. And since survey data is not error-free, a first step in performing a high-quality data fusion is a thorough assessment of the quality of the input files.

Radner *et al* (1980) also describe a fusion application done by The Office of Tax Analysis (OTA) of the USA Treasury Department. The OTA investigated a different method of constrained data fusion to ensure that the weights and distributions are preserved in the fused file. Rather than splitting the weights based on rankings, they used the transportation algorithm to simultaneously minimize the distances between records and allocate the weights accordingly. Kadane (1975) describes the use of the transportation algorithm and its associated restrictions mathematically.

To reduce the number of computations needed the OTA first partitioned the data sets into subsamples and applied the transportation algorithm within each of the subsamples. They subsequently applied this approach in two separate fusion projects, namely a match between the 1973 Statistics of Income and CPS data files, and that

between the 1975 Statistics of Income and 1976 Survey of Income and Education files.

Wolff (1977) applies a method developed by Yale University and National Bureau of Economic Research to construct the Measurement of Economic and Social Performance (MESP) synthetic data file. This method focuses on data fusion procedures for very large data sets and objective selection of the set of common variables. The initial step in the fusion process involves creating a cell code for each record in both the donor and recipient files, reflecting the values of the common variables X.

In order to reduce the number of cell codes, ranges of these codes were identified for which the distributions of variables, not used as part of the common set, were significantly different. These ranges were estimated using Chi-squared tests and correlation coefficients. If the distribution of non-common variables for two adjacent subsets, based on the cell ranges, was not significantly different from one another, the two ranges could be combined. If there was a difference, adjacent ranges in each subset were tested again. The process was repeated until a detailed nested structure was completely defined. Both the donor and recipient files were then sorted based on the nested structure, and one or more donor records were linked with each recipient record in an unconstrained approach. Given the technology at the time, this procedure reduced the computational time when fusing large data sets.

The MESP database was the result of three data fusion applications and consisted of asset, liability and demographic information for approximately 60,000 U.S households. As this database was rich in information it could be used for many different analyses. Wolff (1977) used it specifically to estimate the distribution of household wealth.

The first step in the MESP fusion was to combine the 1969 IRS TM with an augmented version of the 1970 IRS TM. The latter contained race and age that was linked with Social Security Administration data through exact record linkage. Matching was based on percentile ranks within cells to account for any differences of the size of adjusted income due to the fact that the information was recorded for different years (1969 vs. 1970). The synthetic file was then further matched with the 1970 Decennial Census 15% Public Use Survey (PUS). A third synthetic file was created by supplementing the 15% PUS with additional data on stocks of some consumer durables from the 1970 5% PUS, that were not recorded in the 15% PUS. For all three fusion applications, cells for each critical variable were defined and records were linked within cells using a number of matching variables.

Ruggles, Ruggles, Wolff (1977) performed a number of tests on the synthetic file from the third MESP fusion to evaluate the accuracy of the match. Regression analysis was run on the original data as well as the fused data set and the regression coefficients for the corresponding variables were compared using Chow tests. Forty of the forty-two tests showed no difference in the coefficients. The authors conclude that these results were a good indication that the synthetic file can be used as a reliable source for modelling.

Rodgers (1984) notes that data fusion procedures were developed without any theoretical justification and although some efforts were made to address these issues, he expressed concern about the validity of the CIA and the impact of violating this assumption. He also notes that a complete synthetic file is not always the solution to the research objective and that macro approaches may be more appropriate. He further reiterated that the choice of the correct set of X variables is very important to the success of the fusion and to ensure that the CIA is true.

O'Brien (1991) describes the first commercially available fusion in the United Kingdom, namely the Target Group Ratings (TGR) database. This database was the

result of a fusion between the Target Group Index (TGI) and Broadcasters Audience Research Board (BARB) data files. The TGI is a continuous product and media survey, while the BARB data records minute-by-minute television viewing for 3,000 households in the UK. The objective of this fusion was to create a single data source that provided detailed consumer information and behaviour for media planning purposes.

The TGI data file was fused to the BARB data file based on thirteen common variables. The variable *gender* was identified as a critical variable. An importance weight was derived for the remaining twelve common variables, within the levels of *gender*, using analysis of variance. The Mahalanobis distance measure was used and a penalty weight was imposed if the donor was used more than once. After months of exploration and validation, the researchers concluded that they were able to successfully fuse the individual sources. The TGR database was launched in 1991.

Rässler and Fleischer (1998) performed a simulation study and compared the results for trivariate normal and lognormal data, simulated for different sample sizes, and fused using nearest neighbour techniques with different linking algorithms. They state that the fusion was stable only if the CIA holds true. Any deviations from conditional independence resulted in an incorrect representation of the true relationship between the sets of unique variables.

Tchaoussoglou and Van der Noort (1999) describe how data fusion was evaluated as a solution to the problem of questionnaire overload in the SummoScanner database, an ongoing survey of print readership, radio listening, TV viewing and socio-demographic information in Holland. This survey is administered nationally through telephonic interviews. Over time the number of titles or publications for all the different media channels increased and the SummoScanner questionnaire gradually became very long. Together with the decline in the public's willingness to participate in surveys, this had a negative impact on the readership results.

A fusion exercise was performed on an existing SummoScanner database that contained all the information for 32,000 respondents. This file was initially split into two equal groups (Scanner 1 and Scanner 2) with half of the print media titles in each file together with a set of common variables. However, the researchers felt that media behaviour should be included in the set of common variables but this would again increase the interviewing time and essentially defeat the purpose.

To solve this problem they applied a four-way split design for print media. In this design the print titles were divided into four sets: A, B, C, D and the initial two split files were further split based on different combinations of the four sets of print media titles. Four matches were done between files that had one set of print titles in common. This set could then be used as part of the common variables without increasing the questionnaire length.

The fusion was evaluated using the estimated proportion of seven target groups exposed to selected publications. Four of these target groups were not identified using any of the common variables, i.e. the analysts had no control over the accuracy of the matching for these groups. Overall they found that the fused file compared very well with the original database. The results were so encouraging that the SummoScanner client implemented the four-way split design in July 1999. The only concern about this design for the SummoScanner data was from the publishers of special interest magazines with very specific target markets. They felt that the fusion may not accurately account for the readership behaviour of their target market and could lead to incorrect results for those groups.

From the time that Sims (1972) first commented that the CIA must be valid in reality for a fusion to be successful, much of the research about data fusion centered on the validity of the CIA, the impact of violating it, and the existence of possible alternative identifiable models. In 2000, a case study was conducted in order to evaluate the effect of violating the CIA on the quality and accuracy of a fusion (Ingram *et al*,

2000). A synthetic data file was constructed between the CPS and the National Health Interview Survey (NHIS). A third data set, the National Survey of American Families (NSAF), was used to evaluate the validity of conditional independence. This was possible as the NSAF contained some information about health (**Y**) and income (**Z**) as it is measured in the CPS and NHIS files.

The set of appropriate critical variables was first identified using regression analysis and the importance of each variable assigned according to its predictive power. A constrained match was then performed between the two separate sources using predictive mean matching within levels of the critical variables. The 1997 NSAF was then used to compare the true relationship between the unique variables with the relationship established through data fusion using the ratio of the Chi-squared values for the synthetic file vs. the NSAF file. This analysis indicated that only a third of the information about the relationship between health and income was captured in the statistically matched file. They therefore concluded that the CIA was not a valid assumption for this data and the effect of violating this assumption was that the association between unique variables were not preserved in the synthetic file.

As the results were only reported for two variables believed to be strongly related, the authors proposed further research and analyses for health and income variables with moderate as well as weak relationships. They suggested that CIA violations will matter less if the relationship between the unique variables is either moderate or weak.

Soong and De Montigny (2001) presented an application of data fusion in media research in Latin America at the ARF Conference. The objective of the study was to optimize multimedia advertising schedules for specific consumer groups. A synthetic data file was created using the TV People Meter panels (TAM) and the TGI Multimedia and Product Usage survey. A constrained match was performed to link the closest records using the transportation algorithm. The quality of the fusion was

indirectly evaluated using some claimed TV viewing information on the TGI survey as a surrogate for the TAM measures in the form of claimed TGI TV ratings.

The TGI survey was split into two subsamples and fused using the set of common variables X. The TGI TV ratings for the two subsamples were compared with that in the matched file for 54 ratings among 271 demographic and product usage categories in Mexico. A strong correlation between the original and fused ratings indicated that the fused results were close to the original. Based on their results and the assumption that the TGI ratings were surrogates for the recorded TAM ratings, the analysts concluded that the TGI / TAM fusion was successful.

The Department of Works and Pensions in the United Kingdom used data fusion to develop the Pensim2 model that will allow simulation of pension policy scenarios (Redway, 2003). Such a model would need a large data set with numerous variables as the basis for simulations. No single source exists that would include personal and household status, income, historical information about the individual's pension contribution and other key population characteristics. A synthetic data file was therefore created by matching three separate sources on similar characteristics. These sources included the Family Resource Survey, the British Household Panel Study (BHPS) and the Lifetime Labour Market Database, which is a 1% sample of National Insurance records that are linked to historical tax data.

The researchers at the Department of Works and Pensions felt that the CIA may hold true for a subset of the unique variables **Y** and **Z**, but not for all. Furthermore, since no single source exists, it would be difficult to prove whether the CIA is a valid assumption or not. It is left up to the researchers to justify any assumptions they made prior to modelling. For this project is was thought that a fusion between the Family Resource Survey and the Lifetime Labour Market Database may violate the CIA because there are only a few common variables and a broad range of Y and Z variables. The BPHS file was more extensive and could act as a bridge between the

other two files. It was also thought that linking this file to both others was more likely to satisfy the assumption of conditional independence.

A constrained match was proposed for the Pensim2 model through the use of the transportation algorithm. However, this algorithm is computationally very demanding and may significantly impact on the time needed to link millions of records. Since they linked three sources, they had to decide in which order this had to happen. With little guidance in the literature regarding this decision, they first performed a match on the Family Resource Survey and the BPHS. This way they did not create a very large database in the beginning of the process, which would have affected computer resources.

Initially they tested the runtime of the transportation algorithm in SAS® but found that it increased significantly for large samples that require more than 700 matches. The sources used in this project were too big for the transportation algorithm to have worked efficiently. They needed to find an alternative approach, given the large number of records in the data. This initiated the development of their Order of Decreasing Difficulty (ODD) algorithm.

The main idea of the ODD algorithm is to first match records for which a close match would be difficult, such as outliers, and remove these from the source files. This was based on ranking of the distance between observations and a measure of central tendency of the common variables in the recipient files, in decreasing order. The observations with the lowest ranks included the outliers and these were selected first to match with the closest observation in the donor file. Matched records were successively removed from the source files and the process repeated until all records were matched. Although their algorithm was computationally faster than the transportation algorithm for very large samples, the transportation algorithm resulted in closer matches.

After Rubin (1986) first introduced multiple imputation as an approach to data fusion that does not rely on conditional independence, considerable research into this has taken place. Rässler (2002) formalizes an approach that creates multiple imputations under an explicit Bayesian model. Moriarity and Scheuren (2004) describe a regression-based algorithm to fuse data that assesses uncertainty in matching, using values from a range of plausible values for the covariance between Y and Z. The multiple imputation approach to data fusion is still widely researched and further study in this area involves analysis of small data sets, data that are not normally distributed and fusion of categorical variables.

Becker and Collins (2007) report an application of data fusion that is focused on specific analysis using a method called dynamic segmentation fusion. The fusion involved a link between the National Media and Marketing Survey conducted in the U.S. and the NetView database, measuring Internet behaviour for a panel of households. The purpose of this fusion was to be able to evaluate the relationship between media and product behaviour, and Internet consumption.

However, this relationship is complex and Becker and Collins (2007) claim that it cannot be fully determined through a single set of common variables. The solution was to define the optimal set of critical and matching variables for a specific target group, such as Internet users. This was done through regression trees, applied to the NetView data file. Time spent on the Internet per month was used as the dependent variable with a large number of independent variables such as demographic information and all common Internet variables. Respondents were classified into the homogeneous subsets as defined by the tree in both data sources. The separate files were then fused within these subsets using the transportation algorithm to constrain the weights and ensure that all currencies were preserved. No external source of information existed to verify the validity of the fusion and the results could only be evaluated from a logical perspective.

New fusion techniques are constantly evaluated to find models that will improve on existing models such as the non-parametric local linear regression estimator (LLR), introduced by Conti, Marella and Scanu (2008). The effectiveness of this algorithm was compared to that of hot-deck and *k*-NN procedures through a simulation exercise. The marginal distributions were slightly better preserved in the LLR and *k*-NN with random residual approaches and random hot-deck produced slightly better results for the conditional distribution $Z \mid X$.

3.2.2 Summary

Data fusion as a data integration technique has been applied and evaluated for many years and in many different disciplines, such as econometric modelling, policy development and market research. Several of the data fusion projects showed promising results, while others were not so convincing.

From early on, the technique was not without problems. The initial development of data fusion methods did not involve any strong theoretical basis. Sims (1972) was the first to highlight the weaknesses in the assumption of conditional independence, which is a mathematical consequence of the single imputation data fusion approach. This matter continues to be the main concern regarding fusion applications that use the CIA as the underlying model that describes the relationship between variables that were not jointly observed.

Other observations regarding the quality of a fusion centered on the quality of the individual data sources, as well as the choice of the set of common variables to ensure the maximum predictive power. Common to all the fusion applications is the amount of time necessary to fully explore and validate the analysis. In short, there is certainly no quick solution to data fusion.

Recent years have seen a rise in the use of multiple imputation as a way to perform data fusion without the restrictive assumption of conditional independence. Regardless of the technique, any fusion analysis is still dependent on some or other assumption, whether it is CIA or the assumption of multivariate normality. None of these assumptions can be easily verified. There is however an argument that deviations from CIA lead to greater model misspecification than deviations from normality (Scheuren, 2009). Although multiple imputation methods such as regression-based data fusion seems to provide valid results for simple random samples from multivariate normal distributions, questions about complex samples and data from other distributions remain to be answered.

Alter (1974) comments on the results of the Canadian SCF / FEX fusion and states "*With guarded optimism one may wish to say that we are on the right track, but that we have a long way to go*". Nearly four decades later, this statement is still applicable. Despite all the research into developing new and better methods that would produce a valid fusion, there is still no optimal mathematical solution. Nevertheless, data fusion may be the only practical solution to the problem of response bias as a result of questionnaire overload. As such, it is certainly a technique that requires attention and it is the responsibility of statisticians to continue to investigate methods in order to establish the best possible methodology of data fusion, in all its application.

3.3 Binary Data Simulation

3.3.1 Literature review

Bahadur (1961) developed a method to simulate binary data with a specific marginal distribution and correlation structure by determining the conditional probability $P(X_D = 1 | X_{D-1}, \ldots, X_1)$ based on an expression he defined for the joint distribution of a set of D binary variables. This joint distribution would reflect the required marginal and correlation structure, and the conditional probability was then used to generate each binary variable X_D. However, the definition of the joint distribution is complex and made it difficult to simulate the data for a large number of correlated binary variables.

Since Bahadur (1961), several authors have proposed methods for simulating correlated binary data. Their approaches varied greatly in terms of the statistical methodology used as the basis for the modelling. Kanter (1975), McKenzie (1981), Lunn and Davies (1998), and Oman and Zucker (2001) all developed models for simulating correlated binary data according to stationary autoregressive processes. Kang and Jung (2001) use the beta-binomial distribution to simulate the total number of ones (1's) in D binary variables, and then to get random permutations of n ones and $D - n$ zeros, which generates stationary binary data.

Another approach was proposed by Emrich and Piedmonte (1991) based on the transformation of multivariate normal variables to binary variables. This process involves solving a set of non-linear equations through numerical integration. Leisch, Weingessel and Hornik (1998) developed a function in R (bindata package) that is also based on normal random variables. The mean and variance is defined through the marginal distribution and correlation structure of the binary variables. The required correlation matrix is restricted to a matrix for which its corresponding covariance matrix is at least positive semi-definite.

Lee (1993) proposes a method to simulate non-stationary binary data by defining the entire joint distribution through solving a large number of non-linear equations. Gange (1995) uses iterative proportional fitting to get the joint distribution from a contingency table by fitting a log-linear model. Kang and Jung (2001) also describe a method that involved the complete enumeration of the joint distribution for a small number of variables.

Another set of models are based on certain distributional properties of independent random variables. Park, Park and Shin (1996) describe a process of simulating correlated binary variables through the linear combination of *M* independent Poisson random variables. Alosh and Lee (2001) simplified the model proposed by Park *et al* (1996) by using multiplicative properties of independent Bernoulli random variables.

Qaqish (2003) simulates correlated binary data based on the conditional linear family of multivariate Bernoulli distributions. The process is initialized by simulating X_1, a Bernoulli random variable with specified mean. The remaining *D – 1* variables are then simulated from the conditional distribution $P(X_D = 1 | X_{D-1}, \ldots, X_1)$. An approach similar to Qaqish is that of Farrell and Sutradhar (2006), where the conditional probability is defined through the logit link rather than a linear function of all the variables.

3.3.2 Summary

Over the past 40 years the problem of simulating binary data according to a set structure has received considerable attention. Several different authors contributed to the development of such simulation models. These models varied in the computational complexity and flexibility to handle specific requirements for the marginal and correlation structures. There are advantages and disadvantages to all the proposed methods.

Farrell and Rodgers-Stewart (2008) review and compare the different approaches in the literature and highlight the strengths and weaknesses of each model. Some examples of the shortcomings include the failure to deal with a large number of binary variables, and some models being restricted in terms of the correlation matrix and only allowing positive correlations. Algorithms that require numerical integration or solving non-linear equations add to the computational complexity of the models. The models where the complete joint distribution of all configurations of *D* binary variables is specified or derived also become impractical for large *D*. For the method proposed by Qaqish (2003) different permutations of the vector of *D* marginal distributions may lead to different joint distributions of the random variables $X_1,\ldots,X_D$.

In the words of Qaqish (2003), "*No single simulation method is expected to be able to cope with all possible* (μ, R) *and easily handle moderate to large n*". The choice of model to use when simulating correlated binary data depends on the objective of the study. The analyst must consider issues such as the stationarity of the marginal distribution, the range restrictions on correlations (e.g. are negative correlations a necessity), the number of simulated variables required and the computational complexity of the algorithm.

3.4 Synthesis of the Literature Reviews

Data fusion is concerned with unit non-response. However, in survey research, missing data is almost unavoidable and item non-response often occurs when a respondent refuses to answer a particular question. Although this must be examined and dealt with during data preparation, item non-response is not part of the scope of this research report.

Rässler (2002) states that there is division among statisticians about the feasibility of data fusion in practice, specifically regarding the validity of the CIA. This can be

seen from the literature as some authors report acceptable results, such as Ruggles *et al* (1977), O'Brien (1991) and Tchaoussoglou and Van der Noort (1999). In contrast to this, others are not in favour of the technique, most notable Rodgers (1984). The literature provides some conjectures about certain conditions that may influence the success of a fusion under the assumption of conditional independence. For example, Ingram *et al* (2000) suggest that the CIA may be less restrictive if the data have a weak or moderate correlation structure. Redway (2003) states that the CIA may be valid for a subset of the unique variables, but it is not possible to determine the set of variables for which this is true to ensure a valid fusion.

Despite the large number of binary simulation algorithms that are available, there is no single recommended method as there are advantages and disadvantages to all. For this research report it is necessary to generate binary data with a specific marginal distribution and correlation structure, and it can be assumed that the correlations are positive in the market research context. Therefore the algorithm of Alosh and Lee (2001) is the most appropriate.

4 METHODOLOGY

The study consists of three separate phases, namely data simulation, data fusion and fusion evaluation. The first phase is concerned with data preparation where data are simulated to reflect specific structures and assumptions and the binary simulation algorithm is used to determine the joint probability distribution for D binary variables.

In the data fusion phase the micro-level data sets are constructed (generated) using the output from the binary simulation algorithm, each of size n = 2000. For each generated data set the level of conditional independence in the data is quantified, followed by the data fusion parameter estimation. The results are evaluated and interpreted in the final phase. The calculations, models and statistical analyses involved in each phase are discussed in detail in the following sections.

4.1 Data Simulation

4.1.1 Generate input

The binary simulation algorithm proposed by Alosh and Lee (2001) requires the marginal distribution of D binary variables and a positive correlation matrix as input. Since market research data are often positively correlated due to the nature and structure of the survey questions, this algorithm can be used to simulate data that reflect real-world situations in the context of market research. As indicated in section 1.2, this analysis is restricted to four binary variables (D = 4), as this will be sufficient to address the research questions.

For each simulation, a different correlation matrix will be randomly selected from the valid range of correlations for a single pre-specified marginal distribution for the four binary variables. The probability distribution assigned to this marginal vector covers

a range of possibilities rather than a single value for all variables and is chosen such that it will yield only positive correlations between the four binary variables. Furthermore, each input correlation matrix is selected such that it reflects differing strengths of the relationship between the variables Y and **Z**. In order to evaluate the impact of conditional independence on the success of a binary fusion, the input correlation matrices will also reflect varying degrees of conditional independence in the data.

A necessary condition for the binary simulation algorithm is that the input correlation matrix must be positive definite. Since not all square matrices are positive definite, it is quite possible that a randomly selected correlation matrices is singular. The **corpcor** library in R has a built-in function that finds the nearest positive definite matrix of any symmetric input matrix. The function, *make.positive.definite*, is based on the algorithm of Higham (1988) and is applied to each generated correlation matrix to ensure that it is positive definite.

A final condition for the generated correlation matrix is to ensure that it will produce valid results from the binary simulation. The algorithm performs a feasibility check at every stage of the process and can stop the procedure if it is unable to determine the parameters. It is therefore possible that a positive definite correlation matrix consisting of only positive correlations may not produce valid results. Therefore, each generated correlation matrix is put through the algorithm to determine whether the results are valid. If not, the matrix is discarded and replaced by another randomly selected matrix.

Based on the above, all input correlation matrices will be positive definite, will produce valid results from the binary simulation algorithm, will reflect a certain level of strength between the unique variables, and will display some degree of conditional independence.

For this analysis 1,000 input correlation matrices are generated for each of three categories of correlation strength (high, moderate, low), reflecting the absence of CIA. Each matrix is then used to further generate an additional nine matrices that reflect differing levels of CIA. Therefore, a total of 30,000 correlation matrices will be used as input to the binary simulation algorithm, together with a single marginal distribution for four binary variables. The steps involved in this process are discussed in detail below.

The first step in generating the input correlation matrices is to define the range of plausible values for binary variables. Such values depend on the marginal probabilities of pairs of binary variables. In particular, the pairwise probability, p_{ab}, of any two binary variables *a* and *b* is constrained as a function of p_a and p_b, such that

$$\max(p_a + p_b - 1, 0) \le p_{ab} \le \min(p_a, p_b).$$

But p_{ab} can be expressed in terms of the marginal probabilities and the correlation between variables *a* and *b* through the equation

$$p_{ab} = \rho_{ab}\sqrt{p_a q_a p_b q_b} + p_a p_b .$$

It therefore follows that the range of each correlation ρ_{ab} is defined by

$$\max\left(\frac{p_a + p_b - 1 - p_a p_b}{\sqrt{p_a q_a p_b q_b}}, \frac{-p_a p_b}{\sqrt{p_a q_a p_b q_b}}\right) \le \rho_{ab} \le \min\left(\frac{p_a - p_a p_b}{\sqrt{p_a q_a p_b q_b}}, \frac{p_b - p_a p_b}{\sqrt{p_a q_a p_b q_b}}\right). \quad (4.1)$$

Based on the range of the valid correlations, as defined in equation (4.1), three levels of strength are identified for the relationship between the unique variables (ρ_{YZ_1} and ρ_{YZ_2}) namely high, moderate and low correlations. Correlations between the 80th and 95th percentile are classified as high correlations, medium correlations are between

the 65th and 80th percentiles, and low correlations are selected between the 50th and 65th percentile. The correlations ρ_{XY}, ρ_{XZ_1}, ρ_{XZ_2} and $\rho_{Z_1Z_2}$ are fixed as moderate to high. This is to ensure that any changes in the quality of the fusion are not due to changes in the relationship between the variables in the individual data sources.

To select an appropriate marginal distribution, the ranges for high, medium and low correlations were assessed for different marginal probabilities. Such a vector should yield positive correlations for all levels of strength, and should also represent incidence for each binary variable that is realistic for survey-type data. The chosen vector of marginal probabilities is given as

$$P = (p_X, p_Y, p_{Z_1}, p_{Z_2})' = (0.7, 0.6, 0.8, 0.5)'. \quad (4.2)$$

Using the marginal distribution given in equation (4.2), the valid ranges of correlations between the four binary variables are defined based on equation (4.1) and the definitions of high, moderate and low correlations. These are given in Table 4.1. The first part of the table shows the range of valid correlations and the range of moderate to high correlations for the relationships between the common variable (X) and each of the unique variables, and between the two Z variables. In the second part of the table the relationship between the unique variables (Y, Z_1) and (Y, Z_2) are defined for all three levels of strength.

A correlation matrix is randomly selected from the desired correlation strength, based on the ranges defined in Table 4.1, such that conditional independence is absent. This correlation matrix is used as a starting point for generating additional input matrices for differing levels of conditional independence. Let this matrix be referred to as the initial correlation matrix. It is denoted as follows

$$R_N = \left\{ r_{ij}^N \right\} \equiv \text{No CIA level.}$$

		ρ_{XY}	ρ_{XZ_1}	ρ_{XZ_2}	$\rho_{Z_1Z_2}$
Range	Lower	-0.535	-0.327	-0.655	-0.500
	Upper	0.802	0.764	0.655	0.500
Moderate to	Lower	0.334	0.382	0.196	0.150
High	Upper	0.735	0.709	0.589	0.450

		ρ_{YZ_1}	ρ_{YZ_2}
Range	Lower	-0.408	-0.816
	Upper	0.612	0.816
High	Lower	0.408	0.490
	Upper	0.561	0.735
Moderate	Lower	0.255	0.245
	Upper	0.408	0.490
Low	Lower	0.102	0.000
	Upper	0.255	0.245

Table 4.1: Valid range of generated correlation matrices

CIA is said to be absent in the correlation if both the partial correlations $\rho_{YZ_1 \cdot X}$ and $\rho_{YZ_2 \cdot X}$ are greater than 0.05. This is based on the significance of a partial correlation between two variables, say Y and Z_1, while controlling for a third variable X (Weatherburn, 1952), for a sample of size 2000.

To illustrate this, consider a partial correlation of 0.05 between Y and Z_1, given univariate X, for a sample of 2000:

$$H_0 : \rho_{YZ_1 \cdot X} = 0 \qquad \text{vs.} \qquad H_1 : \rho_{YZ_1 \cdot X} \neq 0$$

The test statistic for this hypothesis test is

$$t = r_{YZ_1 \cdot X} \sqrt{\frac{n-k-2}{1-r^2_{YZ_1 \cdot X}}} = 0.05 \sqrt{\frac{2000-1-2}{1-0.0025}} = 2.24 \qquad (4.3)$$

where k = number of conditioning variables, namely $k = 1$, and the statistic follows a t-distribution with $(n - k - 2)$ degrees of freedom.

Thus H_0 will be rejected if $|t| > t_{n-k-2,\frac{\alpha}{2}}$, where $t_{n-k-2,\frac{\alpha}{2}} = t_{2000-1-2,0.025} = 1.96$, when testing at the 5% level of significance for a 2-sided test.

Since the absolute value of the test statistic exceeds the critical value, the null hypothesis is rejected in favour of the alternative, at the 5% level of significance. Therefore, a partial correlation of 0.05 for a sample of 2000 is significantly different from zero.

The second step in this phase of the analysis involves generating input correlation matrices that reflect varying degrees of conditional independence. The CIA states that the unique variables Y and **Z** are independent given the common variable X. In statistical terms this implies that the partial correlation between Y and **Z**, given X is equal to zero. If the CIA is true for a given data set, then the following equation is valid

$$\Sigma_{YZ} = \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ}\,. \tag{4.4}$$

Using equation (4.4), it is possible to enforce conditional independence on R_N to create a second correlation matrix for which conditional independence is true, such that equation (4.4) is true. This will result in zero partial correlations $\rho_{YZ_1 \cdot X}$ and $\rho_{YZ_2 \cdot X}$. Let this correlation matrix be

$$R_C = \left\{ r_{ij}^C \right\} \equiv \text{Complete CIA level.}$$

The specific correlations r_{YZ_1} and r_{YZ_2} in the two matrices R_C and R_N provide the lower and upper limits for different levels of conditional independence, ranging from no CIA to complete CIA. The next step involves generating correlation matrices that

reflect deviations from CIA. In total, ten correlation matrices will be generated with differing levels of CIA: one with no CIA, one with complete CIA and eight in between. Seven of the additional eight are generated through incremental deviations from conditional independence, i.e. from R_C. The eighth correlation matrix is generated by randomly selecting correlations for (Y, Z_1) and (Y, Z_2) from the ranges defined by R_C and R_N.

The increment is calculated as

$$Increment = \left(r_{YZ_i}^N - r_{YZ_i}^C \right) / 8 \quad , i = 1, 2 .$$

Each of the seven matrices is then constructed as follows

$$r_{YZ_i}^m = r_{YZ_i}^C + m \times increment \quad , m = 1, \ldots, 7 \quad , i = 1, 2 .$$

All remaining correlations in each R_m are equal to the corresponding correlations in R_N. All additional generated correlation matrices are evaluated to ensure that they are positive definite and will produce valid results for the binary simulation.

In the event that any of the incremental correlation matrices R_m does not produce valid results for the binary simulation algorithm, one or more additional matrices are generated by randomly selecting correlations from the range defined by $r_{YZ_i}^C$ and $r_{YZ_i}^N$. Such matrices may reflect partial conditional independence, indicating that conditional independence is valid for a subset of the unique variables only, instead of for all the unique variables.

4.1.2 Binary data simulation

The algorithm of Alosh and Lee (2001) requires the specification of both the marginal distribution of each variable and the correlation matrix, consisting of positive correlations only, as discussed in section 4.1.2. It uses the property that the product of *M* independent Bernoulli random variables is also a Bernoulli random variable, and its parameter is the product of the original *M* Bernoulli parameters, therefore if

$$U_m \overset{indep}{\sim} Bernoulli(\beta_m) \quad , m = 1, \ldots, M$$

then

$$X = \prod_{m=1}^{M} U_m \sim Bernoulli\left(\prod_{m=1}^{M} \beta_m \right).$$

Consider a *D*-dimensional random vector $(X_1, \ldots, X_D)'$ with a specific marginal distribution $P = (p_1, \ldots p_D)'$ and correlation structure $R = \{\rho_{ij} : i, j = 1, \ldots, D\}$. The objective of the simulation algorithm is to estimate the parameters of $L = D(D+1)/2$ independent Bernoulli random variables, such that the product of selected subsets of these variables result in the desired set of *D* correlated binary variables. Each of the *D* binary variables, X_d, is therefore constructed as

$$X_d = \prod_{l \in C_d} U_l$$

where

$C_d \subseteq \{1, \ldots L\}$ and $U_l \sim Bernoulli(\beta_l)$.

The expected value of a Bernoulli random variable is the probability that the variable takes on a value of one. Furthermore, the correlation between any two binary variables can be expressed in terms of the respective marginal probabilities as well as the joint probability of the two variables. Both of these statistics can be calculated from subsets of the L independent Bernoulli variables with parameters β_l. Therefore, the value of each Bernoulli parameter β_l is a function of both the marginal

distribution and the correlation structure used as input to the simulation and is given as

$$E(X_d) = P(X_d = 1) = p_d = \prod \beta_{C_d}$$

and

$$\begin{aligned} E(X_i X_j) &= P(X_i = 1, X_j = 1) = p_{ij} \\ &= \prod \beta_{C_i \cup C_j} \\ &= \frac{\prod \beta_{C_i} \cdot \prod \beta_{C_j}}{\prod \beta_{C_i \cap C_j}} \\ &= p_i p_j / \alpha_{ij} \end{aligned}$$

where

$$C_i, C_j \subseteq \{1, \ldots, L\}; \quad i, j = 1, \ldots, D; \quad i \neq j; \quad \alpha_{ij} = \prod \beta_{C_i \cap C_j} .$$

Based on the definition of the correlation between two binary variables, namely

$$\rho_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i p_j q_j}} \quad , \; q_i = 1 - p_i$$

it follows that α_{ij} can be expressed as

$$\alpha_{ij} = \frac{p_i p_j}{\rho_{ij} \sqrt{p_i q_i p_j q_j} + p_i p_j} .$$

The binary simulation algorithm of Alosh and Lee (2001) determines the values of all parameters β_l through iterative computation of α_{ij} and identifies the subsets C_d. A data set consisting of *D* binary variables with the pre-specified marginal distribution and correlation structure can now be simulated using the set of L independent Bernoulli(β_l) random variables. This can be done by generating L Bernoulli(β_l) variables of size *n* and computing the product of selected subsets of the L variables. The output from the simulation can also be used to determine the complete joint probability distribution of *D* binary variables.

The simulation procedure of Alosh and Lee (2001) comprises the following steps:

Step 1 – Initialization:

Let iteration $l = 1$.

Compute the initial, upper-diagonal S-matrix:

$S_1 = [\alpha_{ij}^1] \quad , i, j = 1, \ldots, D$.

Step 2 – Determine the parameters:

Let $T_l = \{\alpha_{ij}^l : 0 < \alpha_{ij}^l < 1; i, j = 1, \ldots, D\}$.

Determine β_l and $\{r, s\}_l$:

$$\beta_l = \alpha_{rs}^l = \max\{\alpha_{ij}^l : \alpha_{ij}^l \in T_l\}.$$

$\{r, s\}_l$ = index (i, j) of β_l.

Check feasibility:

If $\alpha_{rr}^l = 1$ or $\alpha_{ss}^l = 1$, then stop (invalid results), otherwise continue.

Determine the index set A_l:

Let $A_l = \{r, s, i, j : 0 < \alpha_{ij}^l < 1; \alpha_{ij}^l \in S_l\}$.

A_l must have as many elements as possible, and be unique for each iteration.

To find A_l:

Let $A_l^0 = \{r, s\}_l$.

For $v = 1, \ldots, D$:

$$A_l^v = \begin{cases} A_l^{v-1} \cup \{i\} & , \alpha_{ij}^l > 0; \forall j \in \{A_l^{v-1}\} \\ A_l^{v-1} & , otherwise \end{cases}$$

then $A_l = A_l^D$.

Step 3 – Update α_{ij}^{l}:

Create the upper-diagonal matrix $S_{l+1} = \{\alpha_{ij}^{l+1}\}$.

Let $\alpha_{ij}^{l+1} = \alpha_{ij}^{l} / \beta_{l} \quad , \forall \{i, j\} \in A_{l}$.

If all $\alpha_{ij}^{l+1} = 1$, go to Step 4, otherwise let $l = l + 1$ and go to Step 2.

Step 4 – Compute the complete joint probability distribution:

Let $U_{l} \sim Bernoulli(\beta_{l}) \quad , l = 1, \ldots, L$.

Create subsets C_{d} from all index sets A_{l} for which the product of the Bernoulli variables included in the set will yield each of the desired *D* binary variables:

$$C_{d} = \{U_{l} : A_{l} = d; l = 1, \ldots, L\}.$$

The joint probability distribution of *D* binary variables consists of 2^{D} possible configurations of zeros and ones. In order to define the complete distribution, the probability associated with each of these outcomes must be determined. Such probabilities can be written as the intersection of the probabilities associated with *D* events, namely

$$P(configuration) = \bigcap_{d=1}^{D} P\{X_{d} \in (0,1)\}.$$

By simultaneous evaluation of all the sets C_{d}, each of the 2^{D} configurations can be generated as the product of subsets of L Bernoulli variables, through the binary simulation algorithm. Therefore, the intersection of the events that L Bernoulli variables take on either a one or a zero, will result in a particular configuration.

$$\bigcap_{l=1}^{L} \{U_{l} \in (0,1)\}.$$

This can be done by creating a vector of length L for all configurations, except "00...0", where the vector elements are either zero or one, such that the product of the

Bernoulli variables in each set C_d will result in the required configuration. To simplify the computation, the probability of the configuration consisting of zeros only can be computed using the complement rule

$$P(00\ldots0) = 1 - P(all\ others). \tag{4.5}$$

Consider the example from Alosh and Lee (2001) where D = 3. The binary simulation algorithm will determine the parameters for six Bernoulli random variables such that

$$P(U_l = 1) = \beta_l .$$

The algorithm generates the subsets of the Bernoulli variables that will produce the three related binary variables, in this case

$$C_1 = \{1, 2, 3\}$$
$$C_2 = \{1, 4, 5\}$$
$$C_3 = \{1, 2, 4, 6\}.$$

Based on these subsets, each of the three binary variables can be created as the product of selected Bernoulli random variables with certain parameters

$$X_1 = U_1 \times U_2 \times U_3$$
$$X_2 = U_1 \times U_4 \times U_5$$
$$X_3 = U_1 \times U_2 \times U_4 \times U_6 .$$

If all six Bernoulli variables take on the value of one, all three binary variables will be equal to one. Therefore, the probability of getting a configuration "111" is the probability of getting the value of one for each of the six Bernoulli variables. Using the properties of the product of independent Bernoulli random variables, this probability can be calculated as

$$P(\{111\}) = \prod_{l=1}^{6} P(U_l = 1).$$

Similarly, the probability associated with configuration "110" is calculated as the product of the first five Bernoulli variables, each with probability of being equal to one, multiplied by the probability that U_6 is equal to zero

$$P(\{110\}) = P(U_6 = 0)\prod_{l=1}^{5} P(U_l = 1).$$

To find the complete joint probability distribution, all possible scenarios that can yield each configuration must be identified. In some instances, only one vector of zeros and ones across the L Bernoulli variables will result in a particular configuration.

It is however possible that more than one vector will produce the same results. For example, consider the configuration "001" for the example above. The third binary variable is equal to one, therefore all four variables in its set, namely U_1, U_2, U_4 and U_6 must all be equal to one, and variables U_3 and U_5 equal to zero in order to generate this outcome. But if U_3 is equal to one while U_5 is equal to zero, it will also result in the same configuration of "001". Therefore the probability of getting this particular configuration is the combination of the probabilities calculated for these two specific outcomes of six Bernoulli variables.

$$P(\{001\}) = P(U_3 = 0)P(U_5 = 0)\left[\prod_{l=1, l\neq 3,5}^{6} P(U_l = 1)\right] + P(U_5 = 0)\left[\prod_{l=1, l\neq 5}^{6} P(U_l = 1)\right].$$

Once the probabilities for the seven specific configurations are established, the probability associated with the outcome {000} can be found using equation (4.5).

4.2 Data Fusion

4.2.1 Generate the micro-level data

The binary simulation algorithm is used to determine the complete joint probability distribution for D binary variables with a particular marginal distribution and correlation structure. This makes it possible to create a micro-level data set of size n. The joint probability distribution indicates the proportion of any sample that consists of a particular configuration of zeros and ones. By applying these probabilities to a sample size, the number of observations in the sample with the specific outcome over D binary variables is created. Such a data set is then seen as the theoretical data that would have been generated if all items in the questionnaire were administered to all respondents.

For example, consider a data set with four binary variables. The binary simulation algorithm is used to determine the probability distribution for the sixteen possible configurations of zeros and ones. Say the probability that all four variables take on a value of zero is equal to 0.1654. Then, for a sample of size n = 2000, a total of 331 observations will have the configuration "0000", since

$$\begin{aligned} n \times P(\{0000\}) &= 2000 \times 0.1654 \\ &= 330.8 \\ &\approx 331. \end{aligned}$$

Due to the size of the probabilities, the resulting number of observations per configuration is often a decimal value. When generating a micro-level data set, the number of observations for each configuration must be integer values, so the values are rounded to the nearest integer. As a result, the total sample size across all possible configurations may not always be equal to the specified value of n. This will not impact on the results since the sample will be approximately equal to the required sample size.

The generated micro-level data set can now be divided into two subsets of approximately equal size, say subsets A and B. If the generated sample size is an even number, each subset will be of size $n_A = n_B = n/2$. If there is an odd number of observations in the sample, one of the subsets (say A), will be of size $n_A = \text{int}(n/2)+1$, while the other will contain $n_B = \text{int}(n/2)$ of the observations. Variables Z_1 and Z_2 are removed from subset A such that it will include variables X and Y only. On the other hand subset B will include the data for X, Z_1 and Z_2. Therefore, Y and **Z** are not jointly observed in the individual data sources.

For a data fusion application, it is very important that the individual data sources are aligned in terms of both variable and sample units. In other words, both surveys are administered to the sample target population and must therefore represent the distributions in the population. An important consideration in this simulation procedure is to ensure that the two individual subsets, A and B, reflect the original distributions in the complete data set. Thus, the effectiveness of the random subdivision of the original data must be evaluated.

The probability distribution across all configurations in the random subset A will be similar to that in the original data file (referred to as file AB) if the difference between the two distributions is approximately zero. This difference is best defined through the sum of squared deviations (SSD) between the probability distributions in files AB and A across all configurations. Similarly, the differences between files AB and B, as well as files A and B can also be determined with the SSD:

$$SSD_1 = \sum_{c=1}^{16} [P_{AB}(config_c) - P_A(config_c)]^2 .$$

$$SSD_2 = \sum_{c=1}^{16} [P_{AB}(config_c) - P_B(config_c)]^2 .$$

$$SSD_3 = \sum_{c=1}^{16} [P_A(config_c) - P_B(config_c)]^2 .$$

If all three sums of squared deviations are approximately equal to zero, then the two subsets can be seen as representations of the target population. In practical applications, sample weights are often used to align sample units from different data sources to achieve this result.

4.2.2 Quantifying the degree of CIA

In the data simulation phase of the analysis, the input correlation matrices are generated such that they reflect differing levels of conditional independence. These levels are initially classified according to a categorical factor with descriptive levels, based on incremental deviations from conditional independence. A key objective of this analysis is to assign a numerical value to each simulated data set that would indicate the degree of conditional independence in the data. In order to do this, a measure on a continuous scale must be defined such that the minimum value of the scale indicates the presence of CIA in the data, and the measure increases with deviations from CIA.

A function of entropy, CMI, will be used to quantify the levels of CIA. In the following section, the notion of entropy, joint entropy, as well as some quantities that are closely related to entropy, such as conditional entropy, mutual information and CMI are defined. These concepts are introduced for discrete random variables only, since this analysis is focused on binary data, although the definition of entropy for continuous variables was presented by Shannon (1948).

Cover and Thomas (1991) define the entropy of a random variable Y, consisting of J possible outcomes $\{y_1,\ldots,y_J\}$, as a measure that indicates the amount of uncertainty about the variable. It is denoted by $H(Y)$ and is estimated from the probability mass function of Y

$$H(Y) = -\sum_{y} p(y)\log_b p(y) \quad , H(Y) \geq 0 .$$

The base *b* of an entropy measure signifies the scale of the entropy. Shannon's entropy uses the base *b* = 2, and is expressed in terms of bits. The entropy distribution of a random variable reaches its maximum if all the outcomes of the variable are equiprobable, i.e. $p(y_j) = 1/J$. The maximum is given by

$$H(Y) = -\sum_{\forall y} p(y) \log_2 p(y) \tag{4.6}$$
$$= \log_2 J .$$

Therefore, the entropy of a random variable can be increased by increasing the probability of an unlikely outcome at the expense of another more probable outcome.

The entropy of a single random variable can be extended to a pair of random variables, since the joint probability distribution (Y, Z) can be regarded as a single variable with levels defined by the combination of the levels of Y and Z (Cover and Thomas, 1991). This is referred to as joint entropy and measures the amount of uncertainty in the two random variables Y and Z together. It is defined as

$$H(YZ) = -\sum_{\forall y,z} p(y,z) \log_2 p(y,z) .$$

Conditional entropy is used to determine how much uncertainty about Y remains given knowledge of another variable Z (Jakulin and Bratko, 2004). It given by

$$H(Y \mid Z) = -\sum_{\forall y,z} p(y,z) \log_2 p(y \mid z)$$
$$= H(YZ) - H(Z).$$

Using the definition of conditional entropy, it is also possible to condition on a joint distribution. It therefore follows that

$$H(Y \mid ZX) = H(YZX) - H(ZX) .$$

An important property of conditioning is that the entropy of a variable is reduced when knowledge of another variable is given (Cover and Thomas, 1991). This is defined as

$$H(Y \mid X) \leq H(Y). \tag{4.7}$$

Another function of entropy is called the mutual information of two random variables Y and Z. Jakulin and Bratko (2004) state that it measures the amount of information provided by variable Y (or Z) about variable Z (or Y). It is denoted by $I(Y,Z)$ and can be calculated in terms of either Y or Z, both of which will produce the same expression

$$\begin{aligned} I(Y,Z) &= \sum_{\forall y,z} p(y,z) \log_2 \frac{p(y,z)}{p(y)p(z)} \\ &= H(Y) - H(Y \mid Z) \\ &= H(Y) + H(Z) - H(YZ) \end{aligned}$$

or

$$\begin{aligned} I(Y,Z) &= \sum_{\forall y,z} p(y,z) \log_2 \frac{p(y,z)}{p(y)p(z)} \\ &= H(Z) - H(Z \mid Y) \\ &= H(Z) + H(Y) - H(YZ). \end{aligned}$$

According to Jakulin and Bratko (2004), mutual information can be seen as a measure of correlation between variables, such that $I(Y,Z) \geq 0$. This measure is equal to zero if and only if Y and Z are independent

$$I(Y,Z) = 0 \Leftrightarrow P(Y,Z) = P(Y)P(Z).$$

The CMI indicates how Y and Z are related in the context of a third variable X (Jakulin and Bratko, 2004). It measures the reduction in uncertainty about Y (or Z) due to knowledge of Z (or Y), when X is given. It is defined as:

$$\begin{aligned} I(Y,Z \mid X) &= \sum_{\forall y,z,x} p(y,z,x) \log_2 \frac{p(y,z \mid x)}{p(y \mid x)p(z \mid x)} \\ &= H(Y \mid X) + H(Z \mid X) - H(YZ \mid X) \\ &= [H(YX) - H(X)] + [H(ZX) - H(X)] - [H(YZX) - H(X)] \\ &= H(YX) + H(ZX) - H(X) - H(YZX). \end{aligned}$$

The CMI is always zero or positive, i.e. $I(Y,Z \mid X) \geq 0$. If $I(Y,Z \mid X) = 0$ it implies that Y and Z are unrelated, given knowledge of X. Therefore, it can be interpreted that the association between Y and Z is completely explained by X. This corresponds to the definition of conditional independence in the context of data fusion. For this reason, the CMI can be used to quantify the level of conditional independence, where a zero value indicates complete CIA and a positive value indicates deviation from CIA to some degree.

In order to evaluate the numerical meaning of this quantified CIA measure, the range of values that it can take on for a probability distribution based on one or more binary variables must be established. Since $I(Y,Z \mid X) \geq 0$, the upper limit of this measure must be determined.

Equation (4.6) states that the maximum entropy is attained when all the probabilities are equal. However, in the context of market research, the objective is not to reach equilibrium for any probability distribution. In this situation the absolute maximum can never be attained for an existing, given distribution. The valid range of the CMI measure must be defined for the distribution of the specific variable(s), not the maximum of the entropy distribution. Since the CMI can be expressed in terms of conditional entropy, the property given in equation (4.7) can be used to define the

maximum for a given distribution. To derive the maximum, this will be evaluated for both Y and Z.

$$\begin{aligned} I(Y,Z \mid X) &= H(YX) + H(ZX) - H(X) - H(YZX) \\ &= [H(YX) - H(X)] - [H(YZX) - H(ZX)] \\ &= H(Y \mid X) - H(Y \mid ZX). \end{aligned}$$

Since

$$\begin{aligned} & I(Y,Z \mid X) \geq 0 \\ \Rightarrow & I(Y,Z \mid X) = H(Y \mid X) - H(Y \mid ZX) \geq 0 \\ \Rightarrow & H(Y \mid X) \geq H(Y \mid ZX) \\ \Rightarrow & \max\{I(Y,Z \mid X)\} = \max\{H(Y \mid X)\} \end{aligned}$$

but

$$\begin{aligned} & H(Y \mid X) \leq H(Y) \\ \Rightarrow & I(Y,Z \mid X) \leq H(Y). \end{aligned}$$

Similarly, it can be shown that

$$I(Y,Z \mid X) \leq H(Z).$$

Therefore, if $H(Y) \leq H(Z)$, then $I(YZ \mid X) \leq H(Y) \leq H(Z)$. And if $H(Z) \leq H(Y)$, then $I(YZ \mid X) \leq H(Z) \leq H(Y)$. From this result it follows that the maximum CMI is given by

$$I(YZ \mid X) \leq \min\{H(Z), H(Y)\}.$$

Thus, the quantified conditional independence assumption measure can be expressed as a percentile of its valid range. It is denoted as qCIA and is calculated as follows:

$$qCIA = \frac{I(Y,Z \mid X)}{\min\{H(Y), H(Z)\}} \times 100\%. \tag{4.8}$$

4.2.3 Fusion parameter estimation

Files A and B can now be linked together through the common variable X. D'Orazio *et al* (2006) note that the multinomial distribution is a very flexible parametric model for fusing categorical or discrete data. Since this analysis is focused on binary data the macro parametric approach to data fusion can be employed to link subsets A and B. This is because there are a finite number of possible outcomes in the joint distribution of *D* binary variables. It is therefore sufficient to estimate the complete joint probability distribution of the fused data file, rather than creating a synthetic, respondent-level data file, as is the case for continuous data. This is achieved through the maximum likelihood estimators of the various components of the joint distribution under the assumption of conditional independence.

Consider the trivariate multinomial distribution (X, Y, Z) with $I \times J \times K$ categories and parameter vector $\Theta = \theta_{ijk}$, such that

$$\theta_{ijk} = P(X = i, Y = j, Z = k) \quad , i = 1, \ldots, I \quad , j = 1, \ldots, J \quad , k = 1, \ldots, K \qquad (4.9)$$

where

$$\theta_{ijk} \geq 0 \text{ and } \sum_{ijk} \theta_{ijk} = 1.$$

Equation (4.9) represents the probability that a certain configuration of categories can occur across the variables X, Y and Z. This can be applied to the four binary variables in the simulated data, such that variables X and Y each consist of two categories, while the third variable, Z, is the combination of the two binary variables Z_1 and Z_2. Therefore, the possible values of Z in equation (4.9) are {00}, {01}, {10} and {11}. The total distribution then consists of $2 \times 2 \times 4$ categories.

Under the assumption of conditional independence, each component of the joint distribution reduces to the parameters

$$\theta_X = \{\theta_{i..}\},\ \theta_{Y|X} = \left\{\theta_{j|i} = \frac{\theta_{ij.}}{\theta_{i..}}\right\} \text{ and } \theta_{Z|X} = \left\{\theta_{k|i} = \frac{\theta_{i.k}}{\theta_{i..}}\right\}.$$

Therefore, the joint distribution for the multinomial (X, Y, Z) is given by

$$\theta_{ijk} = \theta_{i..}\theta_{j|i}\theta_{k|i} = \frac{\theta_{ij.}\theta_{i.k}}{\theta_{i..}}. \tag{4.10}$$

The MLE for all the required parameters can be computed from the observed marginal and joint distributions of the contingency tables (X, Y) from subset A and (X, Z) from subset B. The estimates are given as follows

$$\hat{\theta}_{i..} = \frac{n^A_{i..} + n^B_{i..}}{n_A + n_B} \quad , i = 1, \ldots, I\,. \tag{4.11}$$

$$\hat{\theta}_{j|i} = \frac{n^A_{ij.}}{n^A_{i..}} \quad , i = 1, \ldots, I \quad , j = 1, \ldots, J\,. \tag{4.12}$$

$$\hat{\theta}_{k|i} = \frac{n^B_{i.k}}{n^B_{i..}} \quad , i = 1, \ldots, I \quad , k = 1, \ldots, K\,. \tag{4.13}$$

The values of all sixteen probabilities associated with the four binary variables are obtained by substituting the estimates from equations (4.11) to (4.13) into equation (4.10), thus creating the estimated complete joint probability distribution for the fused data file.

4.3 Evaluation

The results from each fused data set are compared to the corresponding original data set, addressing all four levels of Rässler's validity assessment procedure. The evaluation process includes determining how many individual records were preserved in the fusion (the hit rate, level 1), a test comparing the estimated correlation structure to the original structure (using the one-sample $\tilde{T}3$-test proposed by Larntz and Perlman in 1985, level 3), and a series of Chi-squared goodness-of-fit tests to compare the realised marginal and joint distributions to the original distributions (levels 2 and 4).

4.3.1 Hit rate of preserved records

In order to assess whether individual values are preserved after the fusion, the *D*-dimensional vector of values for each respondent in the original and the fused data files must be the same. Rässler (2002) refers to this evaluation as the first level of validity. She argues that this is the most difficult to attain, particularly if the data is continuous. For categorical data, this can be assessed by comparing the number of records for each configuration of responses in the fused and original files, which can be expressed as a hit rate for the fusion.

For *D* binary variables there are $C = 2^D$ possible configurations of zeros and ones. These configurations constitute the data for all respondents. If a particular configuration occurred with probability 0.3 in the original data file, but with probability 0.2 in the fused file, one third of the original records for this configuration were not retained in the fused file. The two files correspond with probability 0.2 for this particular vector of zeros and ones. The comparison is done for all L configurations, and the sum of all the probabilities that indicate correspondence will reflect the proportion of the original data file that is recreated in the fused file. This proportion is the hit rate for the fusion.

In mathematical terms, the hit rate is calculated as follows

Let $G^s = (g_1^s, \ldots, g_L^s)'$ and $F^s = (f_1^s, \ldots, f_L^s)'$ be the vectors of the joint probability distribution for each original simulated and fused data file respectively, for simulation $s = 1, \ldots, S$. Then the hit rate of the fusion is given by

$$HR^s = \sum_{c=1}^{C} \min\left(g_c^S, f_c^S\right) \quad , 0 \le HR^s \le 1.$$

4.3.2 $\tilde{T}3$-test for correlation structure

The third level of validity is aimed at evaluating the preservation of the correlation structure of the data. Since the "complete" data is available through simulation, it is possible to compare the correlation structure of the joint distribution (X, Y, Z_1, Z_2) of the original data file with that of the fused data file. The one-sample $\tilde{T}3$-test, proposed by Larntz and Perlman (1985), will be used to test the hypothesis that a sample correlation matrix, calculated from the fused file, is equal to a specific correlation matrix, namely that of the original data set.

Larntz and Perlman (1985) state that the $\tilde{T}3$-test is easy to compute and it performs well for small sample sizes. In contrast to some test statistics that require the sample correlation matrix to be at least positive semi-definite, the $\tilde{T}3$-test will still produce valid results even with singular correlation matrices. This property is a particularly important consideration for the fusion evaluation. In simulating the binary data, the correlation matrices from the generated input are all positive definite. However, it is possible that a random subset of the simulated binary data file may yield a correlation matrix that is not positive definite. This situation does not necessarily hinder the overall analysis and will not be verified. But since it could occur, the $\tilde{T}3$-test is the most appropriate test to use.

Let $R = \{r_{ij}\}$ be the correlation matrix of a *D*-dimensional binary vector for a sample of size *n*, with population correlation matrix $\mathrm{P} = \{\rho_{ij}\}$. The objective is to test the hypothesis H_0: $\mathrm{P} = \mathrm{P}_0$ against the alternative H_1: $\mathrm{P} \neq \mathrm{P}_0$, where $\mathrm{P}_0 = \{\rho_{ij}^0\}$ is a specific correlation structure. Then the $\tilde{T}3$-test statistic is given by

$$\tilde{T}3 = (n-3)^{1/2} \max_{1 \le i < j \le D} \left| z_{ij} - \mu_{ij}^0 \right|$$

where z_{ij} and μ_{ij}^0 are the Fisher z-transform of r_{ij} and ρ_{ij}^0

$$z_{ij} = \tfrac{1}{2} \ln \left[\frac{1+r_{ij}}{1-r_{ij}} \right]$$
$$\mu_{ij}^0 = \tfrac{1}{2} \ln \left[\frac{1+\rho_{ij}^0}{1-\rho_{ij}^0} \right].$$

Therefore, reject H_0 if $\tilde{T}3 > b_\alpha$, where $b_\alpha > 0$ is chosen such that

$$\left[2\Phi(b_\alpha) - 1 \right]^{D(D-1)/2} = 1 - \alpha .$$

Larntz and Perlman (1985) state that this is possibly a conservative level α test for H_0.

4.3.3 Chi-squared goodness-of-fit tests

The final two levels of validity involve an assessment of all marginal and joint distributions in the data. The minimum requirement for any data fusion exercise to be valid is that the marginal distributions, as well as the joint distributions from the separate data sources are preserved in the fused file (level 4). Level 2 is the most important test and deals with the preservation of the joint distribution of variables that were not jointly observed. In practice, this is impossible to test since no information is available to test whether the true joint distribution was retained in the fusion. This level can only be tested through simulation.

From the original data, the overall distribution of X is given by $P_X^{AB} = (p_X, q_X)$, where $p_x = P(X_{AB} = 1)$ and $q_x = P(X_{AB} = 0)$. This serves as the known population distribution of the variable X. The sampling distribution of variable X in the fused file, namely $\hat{P}_X$, can be used to estimate the population distribution P_X of variable X. It is now possible to test the hypothesis that the true population distribution of X, as estimated through the fused data, is equal to P_X^{AB}

$$H_0 : P_X = (p_X, q_X) \qquad \text{vs.} \qquad H_1 : P_X \neq (p_X, q_X)$$

The Chi-squared goodness-of-fit test will be used to compare the marginal and joint distributions in the fused data file with that of the original data file. The frequency distribution for one or more variables in the fused file, written as a vector, is seen as the sample distribution for the variables of interest and can be used to estimate the population parameters. These frequencies are treated as a one-dimensional contingency table.

The function *chisq.test* in R calculates the Chi-squared test for contingency tables, but is also used to perform the Chi-squared goodness-of-fit test for one-way frequencies. The input to the function consists of a vector of frequencies from the sample data, as well as a vector of proportions under the null hypothesis. The test statistic is given by

$$\chi^2 = \sum_{c=1}^{C} \frac{(f_c - e_c)^2}{e_c}$$

where

$C \equiv$ number of classes or levels in the one-way table

$f_c \equiv$ frequency in class c

$e_c \equiv$ expected frequency of class c under the null hypothesis.

Under the null hypothesis the asymptotic distribution of the test statistic is a Chi-squared distribution with (C – 1) degrees of freedom. The null hypothesis is rejected

at the α % level of significance if the value of the test statistic exceeds the critical value of $\chi^2_{\alpha,(C-1)}$. This hypothesis test is done for all marginal and joint distributions.

4.3.4 Output

All levels of validity, except the first, are evaluated using statistical hypothesis testing. Therefore, the success of each test can be measured using the resulting p-value from the hypothesis test. In particular, a non-significant p-value indicates that there is insufficient evidence in favour of the alternative hypothesis. For testing a correlation structure, this implies that the correlation structure from the fused data is not significantly different from that of the original simulated data, used as the population file. Also, non-significant p-values from the Chi-squared tests indicate that the original marginal or joint distributions are retained in the fused file.

On the other hand, a p-value close to zero implies that the alternative hypothesis is possibly true, therefore the specific measures (marginal, joint or correlation) in the fused file do not correspond to that of the original data. The p-values from the hypothesis tests provide a means to quantify the success of a fusion through a numerical measure that has a very particular meaning in statistics.

The hit rate of the fusion (level 1) is the proportion of records from the original data that is retained in the fused file. No statistical tests will be used to determine a "good" hit rate, and it will only be evaluated through descriptive measures. The quantified success of a fusion can now be compared with the qCIA measure.

4.4 Expected Results

Numerous references in the literature discuss the impact of the assumption of conditional independence in data fusion. In particular, it has been shown in many different applications that the CIA must hold true for a fusion to be valid and reflect the true distribution of variables that were never jointly observed. It is therefore

expected that the results will be poor when using data for which the CIA is not true. The main question to be answered is if there is a gradual decline as the data deviate from CIA, or if the deterioration is more abrupt.

Although most of the simulated data sets reflect an incremental deviation from CIA, a number of data sets are also generated by randomly selecting correlations from a specific interval. It is therefore possible that such data sets will reflect a degree of partial conditional independence, in other words, CIA will be true for a subset of the unique variables but not all. It is expected that the joint distribution for those variables for which CIA is true, will be retained in the fused file. However, the overall results are likely to be poor for this situation.

Another component that will be evaluated is based on the conjecture of Ingram *et al* (2000) that the presence of CIA is perhaps less of an issue if the correlations between the unique variables are weak, compared to strong correlations. This will be evaluated to determine if there is any justification for this argument.

5 ANALYSIS

The first section of this chapter (5.1) uses the results from a single simulation to give a detailed discussion of the results obtained from different stages of the simulation and fusion processes, and to show how these results can be evaluated as to the quality of the fusion. Although the results are interpreted in the context of market research, it is important to note that binary data fusion is not restricted to this industry. The overall results for the simulation study are discussed in section 5.2, and the analyses by strength of correlation and for partial CIA are evaluated in sections 5.3 and 5.4 respectively. Section 5.5 reviews all the results in the light of the research questions posed in Chapter 1.

5.1 Single Simulation Analysis

The results for a single simulation could have arisen from a market research survey aimed at collecting information about media consumption and product usage of household purchase decision makers in Gauteng. This survey consists of four binary variables, as defined in Table 5.1. The survey could result in data for the target population, with a sample size of 1999 respondents.

Variable	Description	Codes
X	Gender	1 = Female, 0 = Male
Y	Read Sunday Times	1 = YES, 0 = NO
Z_1	Regularly consume Fanta Orange	1 = YES, 0 = NO
Z_2	Regularly consume Simba potato chips	1 = YES, 0 = NO

Table 5.1: Variable description and code frame

The simulated data set is such that the assumption of conditional independence is valid and there is a moderate to high relationship between variables Y and **Z**. The

required marginal distribution and correlation matrix, used as input to the binary simulation algorithm, are

$$P = (p_X, p_Y, p_{Z_1}, p_{Z_2})' = (0.7, 0.6, 0.8, 0.5)'$$

$$R = \{r_{ij}\} = \begin{pmatrix} 1 & 0.698274 & 0.656700 & 0.483163 \\ & 1 & 0.458556 & 0.337380 \\ & & 1 & 0.430799 \\ & & & 1 \end{pmatrix}$$

5.1.1 Evaluating the simulated data set

For the assumption of conditional independence to be valid, equation (4.4) must be satisfied. This equation can also be expressed in terms of correlations, namely $\rho_{YZ} = \rho_{YX}\rho_{XX}^{-1}\rho_{XZ}$. For this data set, conditional independence is present, since

$$\begin{bmatrix} 0.458556 \\ 0.337380 \end{bmatrix} = [0.698274][1]^{-1}\begin{bmatrix} 0.656700 \\ 0.483163 \end{bmatrix}$$

The correlation between (Y, Z_1) is equal to 0.458556 and between (Y, Z_2) is equal to 0.337380. Based on the ranges defined in Table 4.1, these two correlations are high and moderate, respectively.

The binary simulation algorithm produces the complete joint probability distribution for the four binary variables, which can be used to generate a micro-level data file. Appendix A gives a detailed illustration of the steps in the process.

To ensure that the initial data simulation is done correctly, the input and output of the binary simulation algorithm must be compared, and should be virtually identical. For this example the binary simulation algorithm produced the same marginal distribution and correlation matrix given above (up to ten decimal places).

The data set is created by applying the complete joint probability distribution from the binary simulation algorithm to a sample size of n = 2000. When creating a micro-level data set, the sample size allocated to each of the sixteen possible configurations of zeros and ones across four binary variables must be an integer value. Due to rounding the total sample generated is equal to 1999 rather than 2000. Table 5.2 shows the probability distribution and allocated sample sizes.

Some of the probabilities for the sixteen configurations in Table 5.2 are relatively small. It is therefore possible that some information is lost when generating the micro-level data set and that the resulting marginal distribution and correlation structure for the data may be slightly different to the required structures. What is important for this study is that the difference is minimal. The marginal distribution and correlation structure calculated from the generated data are

$$P = (p_X, p_Y, p_{Z_1}, p_{Z_2})' = (0.69985, 0.60030, 0.80040, 0.49975)'.$$

$$R = \{r_{ij}\} = \begin{pmatrix} 1 & 0.697844 & 0.658766 & 0.484289 \\ & 1 & 0.458685 & 0.337629 \\ & & 1 & 0.431540 \\ & & & 1 \end{pmatrix}.$$

The largest difference between the required and generated marginal distributions is 0.0004002001, while the largest difference between correlations is 0.002066414. Therefore, the generated data structures are approximately the same as the original required structures, implying that the micro-level data reflect the required structure and assumptions.

Configuration	Probability	Sample Size
0000	0.165400	331
0001	0.001003	2
0010	0.075112	150
0011	0.035246	70
0100	0.013888	28
0101	0.000084	0
0110	0.006307	13
0111	0.002959	6
1000	0.001210	2
1001	0.002245	4
1010	0.040919	82
1011	0.078864	158
1100	0.005662	11
1101	0.010507	21
1110	0.191502	383
1111	0.369090	738
TOTAL	1	1999

Table 5.2: Generated probability distribution and sample sizes

5.1.2 Quantifying level of CIA

In the generated data set, the level of conditional independence in the data is quantified using the qCIA measure (see Appendix B for calculations). Using equation (4.8), the qCIA for this example is

$$qCIA = \frac{I(Y, Z \mid X)}{\min\{H(Y), H(Z)\}} \times 100\%$$
$$= 0.015739.$$

The values of the qCIA for the 30,000 simulations are in the interval [0, 60). A qCIA value of 0.015739 is very close the lower bound of possible values and indicates the presence of conditional independence. This can be confirmed by checking the partial correlations. Per definition, if conditional independence is present, then the partial correlations $\rho(YZ_1 \mid X)$ and $\rho(YZ_2 \mid X)$ must be zero, or at least approximately zero. For this example $\rho(YZ_1 \mid X) = -0.00191$ and $\rho(YZ_2 \mid X) = -0.00053$. Based on these partial correlations, it can be assumed that conditional independence is present in this data.

Therefore, for this data set the assumption of conditional independence is true, and the relationship between the variables Y and Z is fairly strong.

5.1.3 Binary data fusion

The next phase in the analysis is concerned with binary data fusion. To achieve this, the generated micro-level data set is divided into two random subsets of approximate equal size, such that one subset contains variables (X, Y) and the other consists of variables (X, **Z**). Let the original generated data set be set AB, and the two subsets be referred to as subsets A and B respectively, with sample sizes n_A = 999 and n_B = 1000.

In the event that the entire survey is administered to a random sample from the target population, the resulting data would be viewed as an estimate of the true population distribution of all the survey data. However, in practical data fusion applications this complete file is not available, and must be estimated through fusing two or more data sets collected from different samples drawn from the same target population. In this illustration, the two separate data sources are subsets A and B. Therefore, the two subsets must be representative samples of the original data set, seen as the true population distribution of the survey data. To see whether this is the case, the SSD

between the joint probability distributions of two data sets are calculated. Data for which the distributions are similar will have SSD values close to zero.

For this data the SSDs are given by

$$SSD_1 = \sum_{c=1}^{16} [P_{AB}(config_c) - P_A(config_c)]^2 = 0.000695.$$

$$SSD_2 = \sum_{c=1}^{16} [P_{AB}(config_c) - P_B(config_c)]^2 = 0.000694.$$

$$SSD_3 = \sum_{c=1}^{16} [P_A(config_c) - P_B(config_c)]^2 = 0.002778.$$

Since these values are close to zero, it indicates that there is not a substantial difference between any two joint probability distributions. Therefore, the two random subsets can be seen as representative samples, drawn from the same target population.

Given that the complete joint probability distribution is easily determined from each subset A and B, this binary data is fused under the macro parametric approach. The MLEs for $\theta_{i..}$, $\theta_{j|i}$ and $\theta_{k|i}$ are used to determine the joint distribution based on two separate data sources, under the assumption of conditional independence.

Tables 5.3 and 5.4 are the contingency tables of the common and unique variables in each random subset. The various components that determine the joint distribution under the assumption of conditional independence are estimated from the data summarized in these two tables. The MLEs for $\theta_{i..}$, $\theta_{j|i}$ and $\theta_{k|i}$ are given in Tables 5.5 to 5.7, based on equations (4.11), (4.12) and (4.13). The various MLE components are multiplied across the different levels of X, Y, Z_1 and Z_2 to produce the complete joint probability distribution, estimated through the fusion and using equation (4.10).

Table 5.8 shows the probability distribution across all sixteen configurations of zeros and ones for the original simulated data and for the estimated fused distribution, namely the MLEs for θ_{ijk}. This table also shows the sample sizes per configuration for a sample of 1999, based on the original and fused probability distributions. Although the number of generated respondents per configuration is not the same in the two samples, the distributions are very similar.

	Y = 0	**Y = 1**	**TOTAL**
X = 0	287	27	314
X = 1	121	564	685
TOTAL	408	591	999

Table 5.3: Contingency table from subset A

	Z_1 = 0		**Z_1 = 1**		**TOTAL**
	Z_2 = 0	**Z_2 = 1**	**Z_2 = 0**	**Z_2 = 1**	
X = 0	170	1	75	40	286
X = 1	8	13	219	474	714
TOTAL	178	14	294	514	1000

Table 5.4: Contingency table from subset B

X = 0	**X = 1**	**TOTAL**
0.300150	0.699850	1

Table 5.5: Maximum likelihood estimates for $\theta_{i..}$

	Y = 0	**Y = 1**	**TOTAL**
X = 0	0.914013	0.085987	1
X = 1	0.176642	0.823358	1

Table 5.6: Maximum likelihood estimates for $\theta_{j|i}$

	$Z_1 = 0$		$Z_1 = 1$		TOTAL
	$Z_2 = 0$	$Z_2 = 1$	$Z_2 = 0$	$Z_2 = 1$	
X = 0	0.594406	0.003497	0.262238	0.139860	1
X = 1	0.011204	0.018207	0.306723	0.663866	1

Table 5.7: Maximum likelihood estimates for $\theta_{k|i}$

Configuration	Generated Distribution		Fused Distribution	
	Probability	Sample Size	Probability	Sample Size
0000	0.165400	331	0.163070	326
0001	0.001003	2	0.000959	2
0010	0.075112	150	0.071943	144
0011	0.035246	70	0.038369	77
0100	0.013888	28	0.015341	31
0101	0.000084	0	0.000090	0
0110	0.006307	13	0.006768	13
0111	0.002959	6	0.003610	7
1000	0.001210	2	0.001385	3
1001	0.002245	4	0.002251	4
1010	0.040919	82	0.037918	76
1011	0.078864	158	0.082069	164
1100	0.005662	11	0.006456	13
1101	0.010507	21	0.010492	21
1110	0.191502	383	0.176742	353
1111	0.369090	738	0.382537	765
TOTAL	1	1999	1	1999

Table 5.8: Generated and fused probability distribution and sample sizes

5.1.4 Fusion evaluation

The success of the fusion is evaluated through a series of descriptive measures and statistical tests. Various data structures from the fused data, as estimated through the MLE procedure described above, are compared to the original generated data set AB, addressing Rässler's four levels of validity through Chi-squared tests, the $\tilde{T}3$ statistics and the fusion hit rate. The fusion hit rate for this example is 0.976381. Therefore, approximately 98% of the original records are retained in the fusion.

The Chi-squared goodness-of-fit test is used to test whether the fused distributions are the same as the original simulated distributions. This is done for the four marginal distributions X, Y, Z_1 and Z_2, as well as the joint distributions (X, Y), (X, Z_1), (X, Z_2), (X, **Z**), (Y, Z_1), (Y, Z_2) and (Y, **Z**). The joint distribution of (Z_1, Z_2) is not considered as the marginal distribution of each Z variables is evaluated individually. The $\tilde{T}3$ test is performed to compare the correlation structure of the fused data file with that of the original data file.

The results from the different hypothesis tests are summarized in Table 5.9, showing the resulting p-value of the test with an indication of the level of significance. At the 5% level of significance, none of the hypothesis tests are significant. Only the marginal distribution for Z_2 from the fused data differs from the original distribution generated in set AB at the 10% level of significance.

Overall, the various structures are retained in the fused data at the 5% level of significance and the fusion is considered to be a success.

Distribution		p-value	Significant
Marginal	X	1	No
	Y	0.874118	No
	Z_1	0.960355	No
	Z_2	0.065111	@10%
Joint	XY	0.920411	No
	XZ_1	0.959390	No
	XZ_2	0.216490	No
	XZ	0.590962	No
	YZ_1	0.936633	No
	YZ_2	0.301667	No
	YZ	0.636727	No
Correlation		0.834567	No

Table 5.9: Comparison of original and fused distributions and correlation structures

5.1.5 Practical interpretation

From the results described above, it appears that the fused data is a very accurate representation of the generated data set AB. In a practical setting, this implies that an analyst will reach the same conclusions when using the fused data and when using the complete data.

Consider the following research questions regarding media consumption and product usage for the survey defined in Table 5.1:

1. Is there an association between reading the Sunday Times and consuming Fanta Orange and Simba potato chips?
2. What proportion of the target population reads the Sunday Times and regularly consumes both products?
3. What is the distribution of media consumption and individual product usage in the market?

4. Is the Sunday Times a suitable channel for advertising special purchases on competitor products, i.e. will the advertisement reach the users of both product categories?

The Chi-squared test of independence can be used to address the first question, where the null hypothesis states that there is no association between reading the Sunday Times and consuming both Fanta Orange and Simba potato chips, and the alternative hypothesis claims that there is an association. When this hypothesis is tested, no significant difference is found.

These results, however, do not necessarily indicate that the actual distribution of media consumption and product usage for generated and fused data are the same. While a statistician would investigate this further through statistical models such as log-linear analysis or ratio estimates, practitioners in the market research industry would probably do this using a descriptive evaluation of the cell values in the media-product contingency table, or with graphical displays.

For this data set, for which CIA holds, Figure 5.1 shows the joint distribution of Sunday Times and Fanta Orange consumption, for both generated and fused data sets. Similarly, the joint distribution of Sunday Times and Simba potato chips for generated vs. fused data is given in Figure 5.2. In both figures, the generated and fused distributions are very similar and the same conclusion will be reached with both data sets.

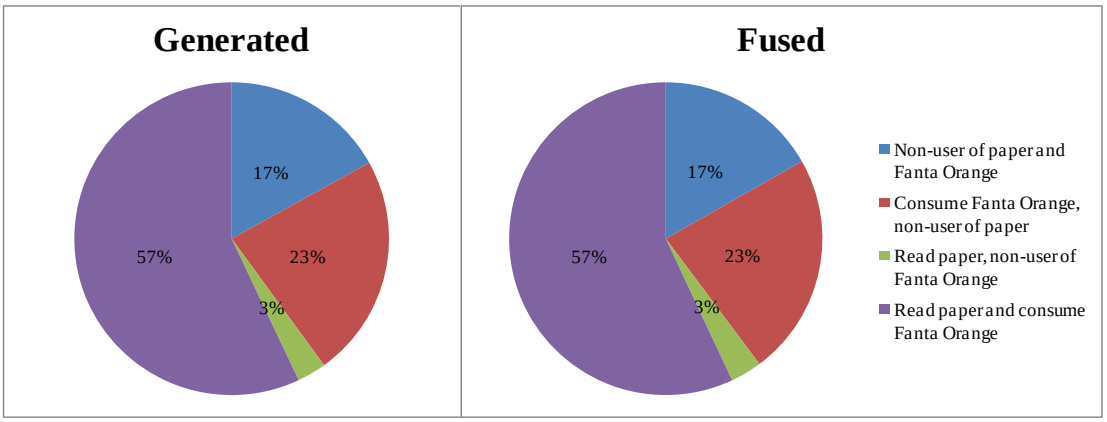

Figure 5.1: Distribution of Sunday Times and Fanta Orange consumption (CIA)

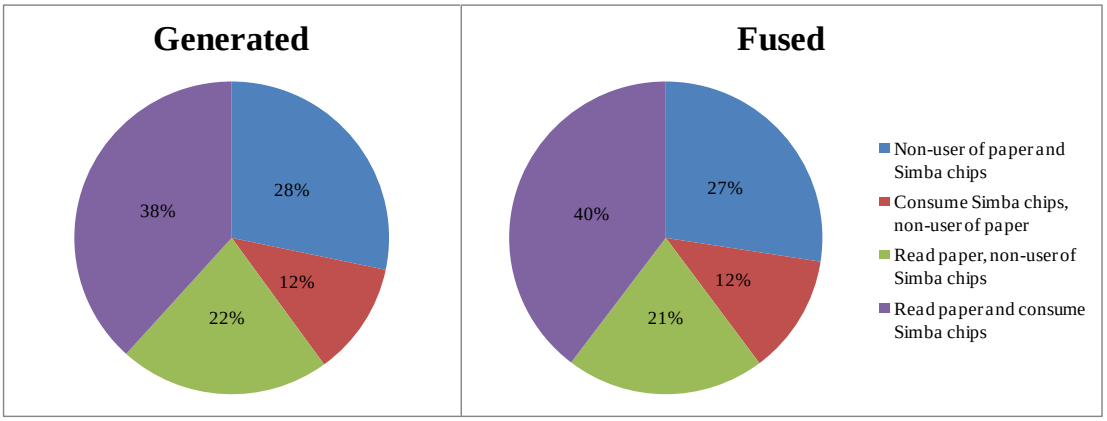

Figure 5.2: Distribution of Sunday Times and Simba chips consumption (CIA)

The assessment of the percentage of the sample that reads the Sunday Times and also consumes both products (Fanta Orange and Simba potato chips) could be done by comparing the confidence intervals for these percentages. For the generated data set, this is (35.1, 39.3), while for the fused data set this is (36.5, 40.7). As these confidence intervals overlap, this indicates that the conclusions would be the same for both data sets. The p-value for the test of equality of two proportions is 0.3616, confirming this conclusion.

To answer the question about the suitability of the Sunday Times as a channel for advertising, the risk ratio (RR) measure, also known as relative risk, can be used to

determine if people who read the Sunday Times are more likely to use both products than people who do not read the Sunday Times. The risk ratio of the joint product usage is the ratio of usage incidence in media consumers to usage incidence in non-users of the media channel.

Dawson-Saunders and Trapp (1994) give the RR calculation and confidence interval for a 2x2 contingency table. For this example the contingency tables is constructed as follows

		Use both products		
		YES	NO	TOTAL
Read Sunday	YES	a	b	a + b
Times	NO	c	d	c + d

Table 5.10: Structure for risk ratio calculation

Based on this contingency table the RR is calculated as

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{a(c+d)}{c(a+b)}$$

and the $(1-\alpha)\%$ confidence interval for the RR is

$$\exp\left(\ln(RR) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1-a/(a+b)}{a} + \frac{1-c/(c+d)}{c}} \right).$$

The RR for the generated data set is 2.17 with associated 95% confidence interval (1.93, 2.45), and that of the fused data set are 2.12 and (1.89, 2.37). Again, the confidence intervals overlap, therefore a market researcher will reach the same conclusion for both the original and fused data sets, namely that Sunday Times

readers are approximately twice as likely to use both products as non-readers of the paper.

For the illustrative example, the conclusions drawn from these analyses will be the same for both original and fused data, since the fused data set reflects the distributions of the generated data set. This is due to the fact that the assumption of conditional independence is a valid assumption. The question is, if the assumption of conditional independence is not completely valid, what impact does this have on the analyses and conclusion drawn from the fused data?

To test this, consider an additional set of eight data sets, each reflecting incremental deviations from the conditional independence established in the example data set, as discussed in section 4.1.1. The research questions posed above are addressed in both generated and fused data sets for various levels of conditional independence (eighteen data sets in total). The aim of this illustration is to summarize the findings from the fused data and compare that to the corresponding findings from the generated data.

The Chi-squared test of independence is applied to all eighteen data sets and all the p-values are less than 0.01, therefore the null hypothesis is rejected for both generated and fused data. Hence, the researcher would reach the exact same conclusions regardless of whether the data was collected using the complete survey, or whether the complete data was estimated through data fusion. Even if the data display deviation from conditional independence, the conclusions are the same.

Graphical representation of the joint distribution of two categorical variables provides insight into the behaviour of consumers with respect to the combination of such variables. This was illustrated in Figures 5.1 and 5.2 for data for which the CIA is true. Figures 5.3 and 5.4 show the joint distribution of Sunday Times and Fanta Orange, and Sunday Times and Simba potato chips, for the generated and fused data in the absence of conditional independence (+8).

From Figure 5.3, one may easily conclude that the joint distribution of media and product usage are completely retained in the fusion, despite the fact that the assumption of conditional independence is not true. However, inspection of the joint media and product usage for the second product category (Figure 5.4) shows that the joint distribution from the fused data set differs from the original generated data. In practical situations the complete data (AB) is not available, so there is no way to establish which subset of variables will produce accurate results through data fusion.

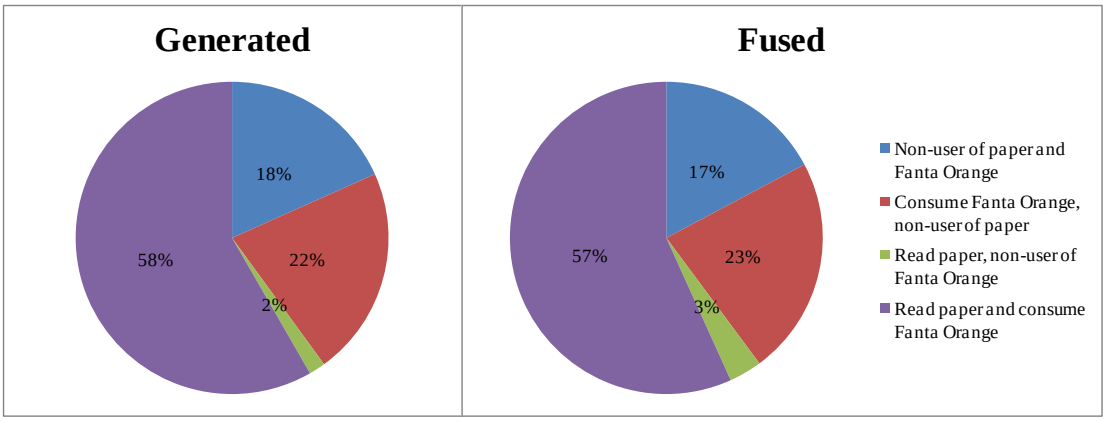

Figure 5.3: Distribution of Sunday Times and Fanta Orange consumption (+8)

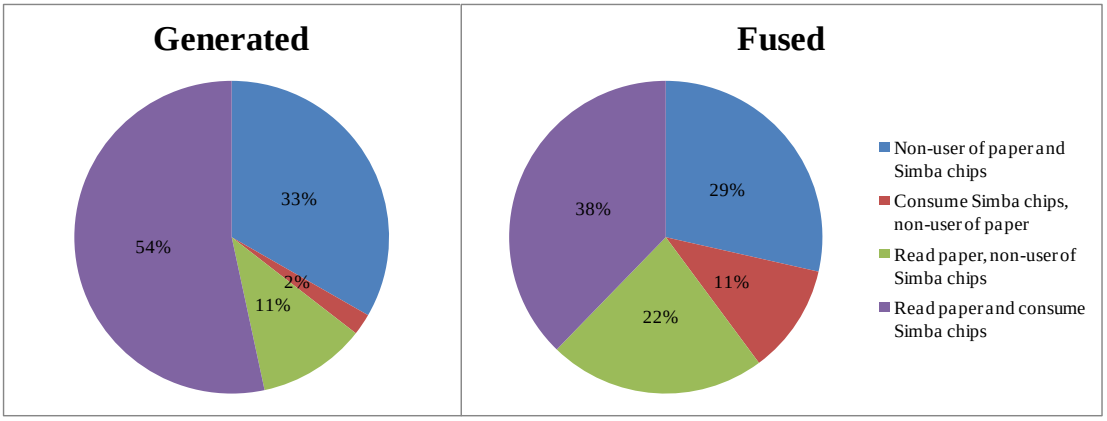

Figure 5.4: Distribution of Sunday Times and Simba chips consumption (+8)

Figure 5.5 shows the percentage of the sample that reads the Sunday Times and also consumes both products (Fanta Orange and Simba potato chips), for data that reflects

differing levels of conditional independence. If each sample is considered to be a representation of the population, then such sample percentages are used to generalize to the entire target population. The 95% confidence intervals for the joint usage percentages are given in Table 5.11.

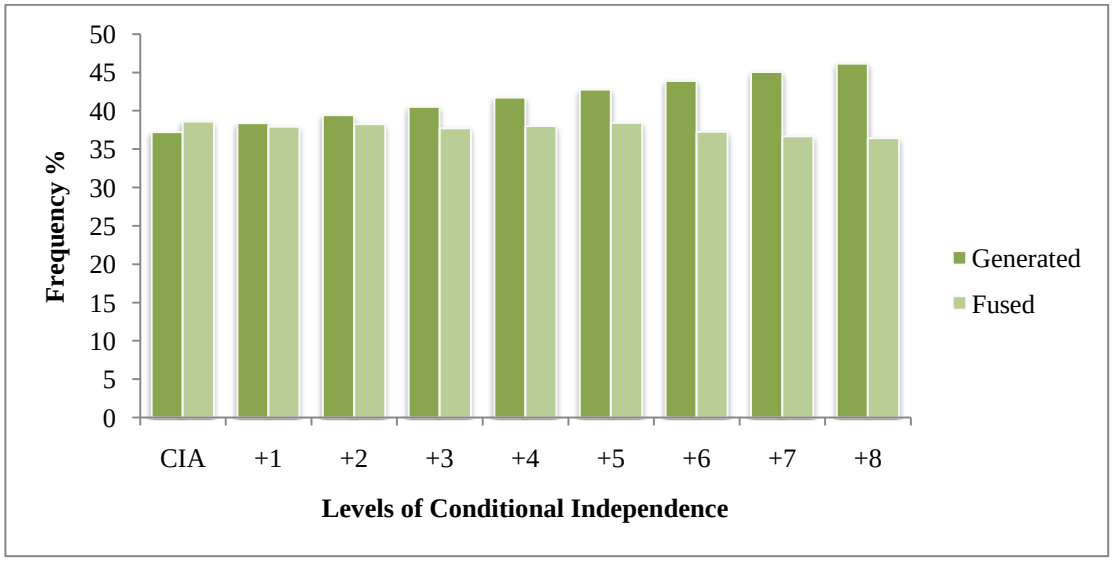

Figure 5.5: Percentages consuming media and both products

	Generated	Fused	Interval Overlap
CIA	(35.1, 39.3)	(36.5, 40.7)	Yes
+1	(36.2, 40.5)	(35.8, 40.1)	Yes
+2	(37.3, 41.6)	(36.1, 40.4)	Yes
+3	(38.4, 42.7)	(35.6, 39.9)	Yes
+4	(39.6, 43.9)	(35.9, 40.1)	Yes
+5	(40.6, 44.9)	(36.3, 40.5)	No
+6	(41.7, 46.1)	(35.1, 39.4)	No
+7	(42.9, 47.2)	(34.6, 38.8)	No
+8	(44.0, 48.3)	(34.3, 38.6)	No

Table 5.11: 95% Confidence intervals for % consuming media and both products

The joint usage of all three categories is similar for the generated and fused data, if the assumption of conditional independence is valid. As the data deviate incrementally from conditional independence, the discrepancy between the generated and fused files becomes more apparent. From +5 incremental deviations, the confidence intervals from the generated and fused data no longer overlap. The two joint usage estimates are significantly different at the 5% level of significance, for all data that deviated substantially from conditional independence. The data that are furthest away from CIA (+8), show a gap of nearly 10% in the sample estimates of the joint usage (46.2% vs. 36.3%). This leads to very different conclusions about the media and product consumption in the population.

Figure 5.6 shows the calculated RR for generate and fused data, for different levels of conditional independence, and the 95% confidence intervals for the RR are given in Table 5.12.

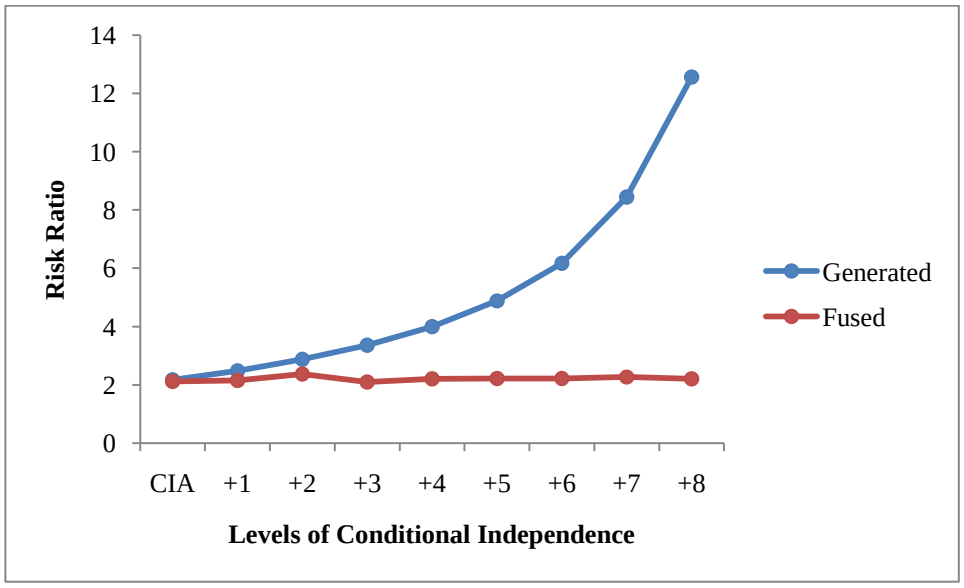

Figure 5.6: Risk ratios of product usage given media consumption

	Generated	Fused	Interval Overlap
CIA	(1.93, 2.45)	(1.89, 2.37)	Yes
+1	(2.19, 2.82)	(1.91, 2.42)	Yes
+2	(2.51, 3.29)	(2.09, 2.70)	Yes
+3	(2.91, 3.88)	(1.88, 2.34)	No
+4	(3.42, 4.67)	(1.96, 2.48)	No
+5	(4.11, 5.79)	(1.96, 2.49)	No
+6	(5.09, 7.47)	(1.97, 2.50)	No
+7	(6.75, 10.57)	(2.00, 2.57)	No
+8	(9.56, 16.50)	(1.95, 2.49)	No

Table 5.12: 95% Confidence intervals for risk ratio estimates

The results follow a similar pattern to those from Figure 5.5. When conditional independence is true, the conclusions from the fused data are correct. But deviations from this assumption lead to differences in such conclusions. From the third incremental deviation, the conclusions drawn from the results are significantly different between the generated and fused data (Table 5.12).

Consider the results for the +8 data set in Figure 5.6. If all the data were collected in a single sample, the researcher would conclude that Sunday Times readers are 12.56 times more likely to drink Fanta Orange and eat Simba potato chips than those who do not read the paper. In contrast to this, when fusing two separate sample, this measure would be equal to 2.21. Based on the latter finding, the market researcher might recommend that this newspaper is not necessarily the best channel for advertisement. It is clear that this marketing decision would be incorrect.

The results above show that, if a data fusion is performed and the CIA is not a valid assumption, any analysis based on the fused data will cause the market researcher to draw conclusions that could be very different to those drawn if the original data set was used.

5.2 Overall Simulation Analysis

The results from the analysis in section 5.1 provide an indication of the success of a binary data fusion for a single simulation. However, significant (or insignificant) results could occur purely by chance. Therefore, a single simulation would not be sufficient to investigate the performance of binary data fusion under the assumption of conditional independence and the analysis is repeated for 30,000 simulations.

5.2.1 Initial evaluation of the simulated data sets

Initial inspection of the 30,000 simulations includes a comparison between the required data structure and the output of the binary simulation algorithm, namely the simulated structures. The second comparison is between the simulated and generated data structures. The minimum and maximum values of the differences between the required, simulated and generated data structures are given in Table 5.13.

		Simulated – Required	**Simulated – Generated**
Marginal distribution	Minimum	0	-0.001298701
	Maximum	0	0.001301301
Correlation matrix	Minimum	0	-0.003526158
	Maximum	0	0.003382289

Table 5.13: Differences between required, simulated and generated structures

In every simulation, the binary simulation algorithm estimated the joint probability distribution such that the marginal distribution and correlation structures are what were required. There are small differences between the simulated structures and those calculated from the generated micro-level data sets. This is primarily due to rounding of sample sizes allocated to a particular configuration of zeros and ones. These differences are very small, indicating that all 30,000 data sets were generated

according to the requirements, and thus reflecting the desired marginal distribution and correlation structure.

The initial correlation matrices are selected from the valid range based on the significant absence of conditional independence, as well as the strength of the relationship between variables Y and **Z**, namely high, moderate and low. Subsequent matrices are generated to reflect differing degrees of conditional independence. As a result, the correlations between Y and **Z** may change from one level of strength to another. The strength of the relationship is ultimately assigned to a simulation based on the correlation matrix of the corresponding generated data set (AB).

The correlations between the unique variables are classified according to the valid ranges defined in Table 4.1. Table 5.14 shows the cross-tabulation of the original three levels of correlation for (Y, Z_1) and (Y, Z_2). Two levels are identified based on this table, namely "strong" and "weak". A strong correlation between the unique variables occurs when both correlations are moderate to high, shown in blue in Table 5.14. Data sets for which the correlations are either both low, or a combination of low and moderate, are classified as having a weak correlation structure between variables Y and **Z** (shown in red). A total of 15,562 simulations are classified as strong and 14,361 as weak correlations. A small subset of the simulations, shown in green, displays a mixture of low and high correlations, 77 in total.

		Correlation YZ_2		
		High	Moderate	Low
Correlation	High	**3999**	**879**	**44**
YZ_1	Moderate	**397**	**10287**	**1855**
	Low	**33**	**692**	**11814**

Table 5.14: Levels of correlation between variables Y and **Z**

5.2.2 Quantifying the level of CIA

The quantified measure of conditional independence of each generated data set is calculated using equation (4.8). Recall that the required correlation matrices were created by first selecting a valid matrix such that the partial correlations are significantly different from zero, and then using the same matrix to enforce complete conditional independence. A total of eight additional matrices were then created to indicate deviation from conditional independence, either through incremental deviation or through random selection from the range defined.

Although the data structures of the generated data set are slightly different from the original required data structures, it is still possible to compare this quantified level of CIA with the ten categorized levels of conditional independence. The lowest level indicates the presence of conditional independence. The following eight levels denote incremental deviations from conditional independence, where the eighth deviation reflects a significant absence of CIA. The correlations for the final CIA level were randomly selected between the correlations for the two extreme CIA levels. This is referred to as the random or mixed. The comparison is shown in Figure 5.7.

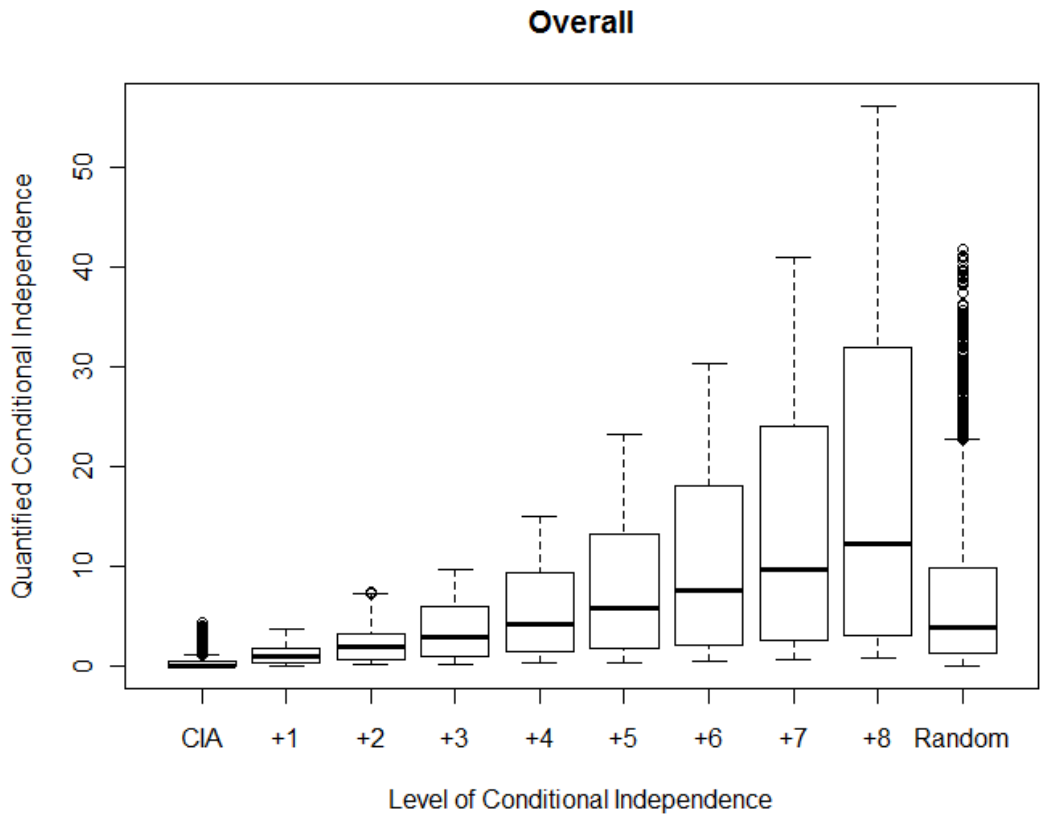

Figure 5.7: Box-and-whisker-plot of the quantified CIA per CIA category

From this graph, the qCIA is equal to zero for the majority of simulations for which the assumption of conditional independence is valid. For incremental deviations away from CIA, the median and range of the qCIA increase. This indicates that the qCIA captures the presence or absence from conditional independence.

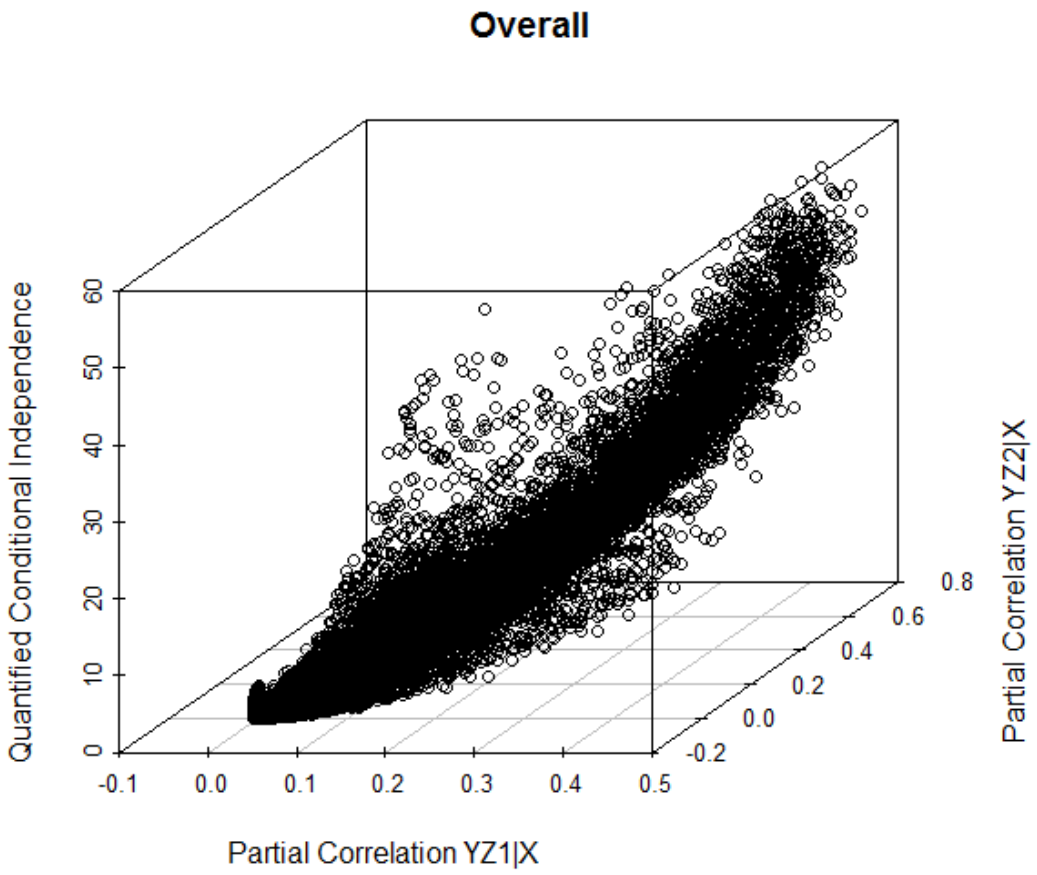

Figure 5.8: Scatter-plot of quantified CIA by partial correlations

The suitability of the qCIA as a measure of conditional independence is further confirmed by comparing the qCIA with the two partial correlations $\rho(YZ_1 \mid X)$ and $\rho(YZ_2 \mid X)$, illustrated in Figure 5.8. This graph shows that both partial correlations are close to zero for low values of the qCIA, and when at least one of the two partial correlations deviate from zero, then the qCIA also deviate from zero. Therefore, this measure can be used to effectively quantify the level of conditional independence present in the generated data.

5.2.3 Binary data fusion

All simulated data sets are randomly divided into subsets A and B, and fused using MLE to create the generated data set AB. Table 5.15 and Figure 5.9 show the SSD

between the generated data set (AB) and the two random subsets A and B. The values given in Table 5.15 indicate that the differences between the joint probability distributions of two data sets are minimal. Visual inspection of all three SSD distributions also indicates that the deviations are small (Figure 5.9). Although there are a number of outliers for the SSD between the two subsets A and B, the majority of these deviations are very close to zero. It can therefore be assumed that subsets A and B are representations of the same target population.

	Minimum	**Maximum**
SSD_1: Set AB – Subset A	0.000035	0.002499
SSD_2: Set AB – Subset B	0.000035	0.002494
SSD_3: Subset A – Subset B	0.000138	0.009988

Table 5.15: Sum of squared deviations summary

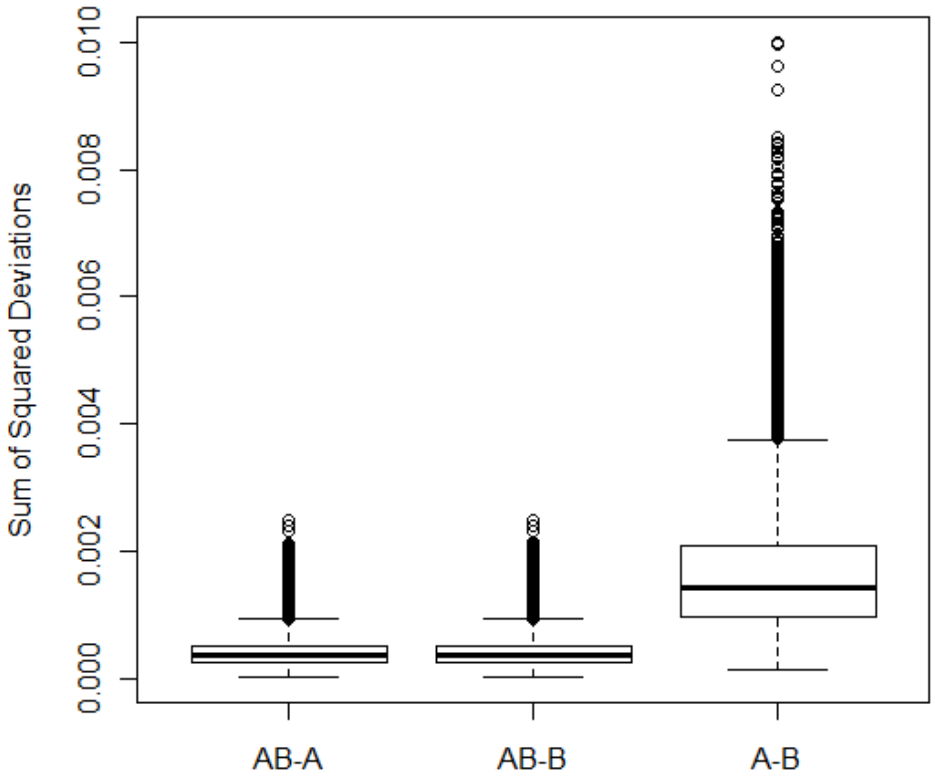

Figure 5.9: Box-and-whisker-plot of sum of squared deviations

5.2.4 Fusion evaluation

The fusion is evaluated using Rässler's four levels of validity. The fourth level of validity is the easiest to achieve as it is concerned with the distributions in the separate data sources. In this section, the fusion is evaluated for all four levels, from the easiest (level 4) to the most difficult (level 1).

Level 4: Preserving marginal distributions

According to Rässler (2002), a minimum requirement for a successful fusion is that the original marginal and joint distributions from the separate data sources are retained in the fused file. It is expected that, regardless of the level of CIA, the distributions will be retained. Chi-squared goodness-of-fit tests are used to compare the marginal and joint distributions from the fused data set with that in the original generated data set AB. If a distribution of a variable in the fused set differs significantly from the original distribution, then the p-value will be significant.

The resulting p-values from all the Chi-squared goodness-of-fit tests for all 30,000 simulations are categorized into four groups to indicate significance. Cat_1 shows the percentage of p-values that were significant at the 1% level, Cat_2 denote p-values between 1-5%, and those that are significant between the 5-10% levels are grouped in Cat_3. Non-significant p-values fall in Cat_4. The percentages within each of these significance categories are given in Table 5.16.

	Categorized p-value			
	Cat_1 [0, 0.01]	Cat_2 (0.01, 0.05]	Cat_3 (0.05, 0.1]	Cat_4 (0.1, 1]
X	0.0	0.0	0.0	100.0
Y	0.4	2.4	3.4	93.8
Z_1	0.3	2.1	3.4	94.2
Z_2	0.6	3.0	4.0	92.4
XY	0.3	1.6	2.3	95.7
XZ	0.5	2.2	3.2	94.0
XZ1	0.3	1.7	2.4	95.6
XZ2	0.4	1.7	2.4	95.6

Table 5.16: Percentages within significance categories

For well over 90% of the 30,000 simulations, the marginal and joint distributions are retained in the fused files. It may appear that there are some anomalies in the results, since a portion of the simulations yielded results that are significantly different from the original generated data. However, in any statistical analysis there is always a chance of error over repeated sample. In the case of repeated testing of samples from the same underlying distribution, a significant result is expected for approximately α% of all analyses, for hypothesis tests at the α% level of significance, and when the null hypothesis is true.

For this analysis, less than 0.5% of the simulations are significant at the 1% level, about 2% at the 5% level, and approximately 5% at the 10% level. The null hypothesis is incorrectly rejected for fewer simulations than expected. This could be as a result of the random division of the original data file into two subsets. If these two subsets are not proper representations of the original data, it implies that the "random sample" does not reflect the true population distributions.

Table 5.17 shows the average of the largest SSD between the original data and the random subsets. The maximum SSD value for all cells in this table is 0.00111.

	Categorized p-value			
	Cat_1	Cat_2	Cat_3	Cat_4
X	-	-	-	0.00041
Y	0.00091	0.00066	0.00056	0.00040
Z_1	0.00074	0.00058	0.00051	0.00040
Z_2	0.00111	0.00075	0.00063	0.00039
XY	0.00091	0.00074	0.00062	0.00040
XZ	0.00089	0.00074	0.00064	0.00039
XZ1	0.00079	0.00059	0.00056	0.00040
XZ2	0.00108	0.00081	0.00068	0.00040

Table 5.17: Largest average SSDs within significance categories

Although the SSD is generally low and close to zero, Table 5.17 shows that the largest average SSDs (AB vs. A or AB vs. B) are consistently higher in the significant categories compared to the non-significant category.

The average qCIA values given in Table 5.18 provide insight into the relationship between the significance of these Chi-squared goodness-of-fit tests and the quantified level of conditional independence.

	Categorized p-value			
	Cat_1	Cat_2	Cat_3	Cat_4
X	-	-	-	6.67
Y	7.77	6.68	6.37	6.67
Z_1	5.38	6.11	6.77	6.68
Z_2	6.56	5.98	6.49	6.70
XY	7.49	6.85	6.67	6.66
XZ	6.99	6.85	6.77	6.66
XZ1	6.48	6.44	6.77	6.67
XZ2	7.96	6.71	5.94	6.68

Table 5.18: Largest average qCIA within significance categories

There is no particular pattern in the average qCIA values across the significance categories. Furthermore, the ranges of the qCIA values for all cells in the above table are from 0 to at least 41.66. This implies that there is no relationship between the Chi-squared output and the level of conditional independence for any of these distributions.

Overall, these results indicate that the minimum requirement for a successful fusion is satisfied.

Level 3: Preserving correlation structures

The $\tilde{T}3$ test statistic is used to assess whether the correlation structure was retained in the fusion. As with the Chi-squared goodness-of-fit tests, the p-values of the $\tilde{T}3$

hypothesis test indicate the success of the fusion. The qCIA is plotted against the p-values of the $\tilde{T}3$ hypothesis tests (Figure 5.10).

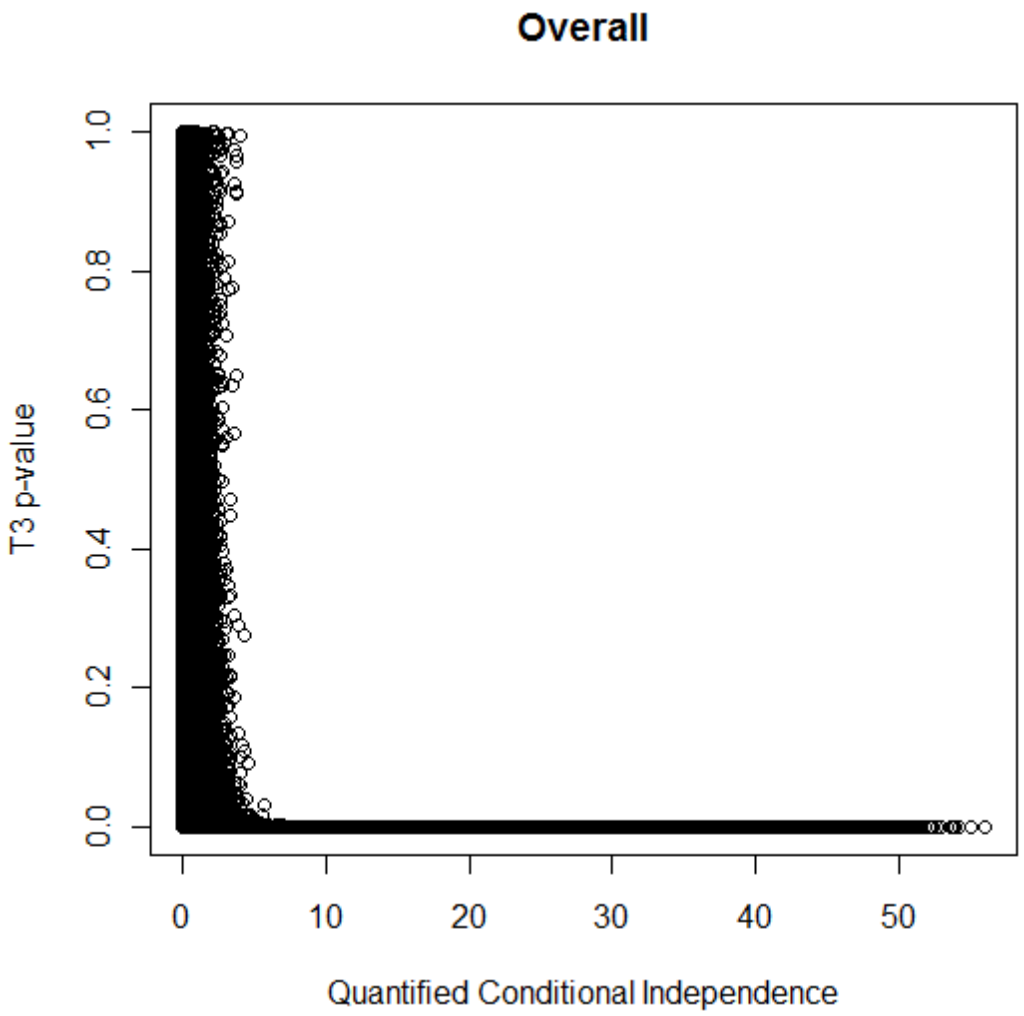

Figure 5.10: Scatter-plot of quantified CIA by $\tilde{T}3$ p-values

This scatter-plot shows that the quality of the fusion, in terms of the correlation structure, deteriorates very quickly with deviations from conditional independence. It is however difficult to see at which level of CIA this happens. This can be further investigated using a visual representation of a contingency table for the two variables, namely the mosaic-plot. The p-values are again grouped into four levels of significance (A, B, C and D), and the qCIA measure is categorized into groups to

identify a certain level of conditional independence, namely [0, 1], (1,2], (2,3], (3,4], (4,5] and (more than 5).

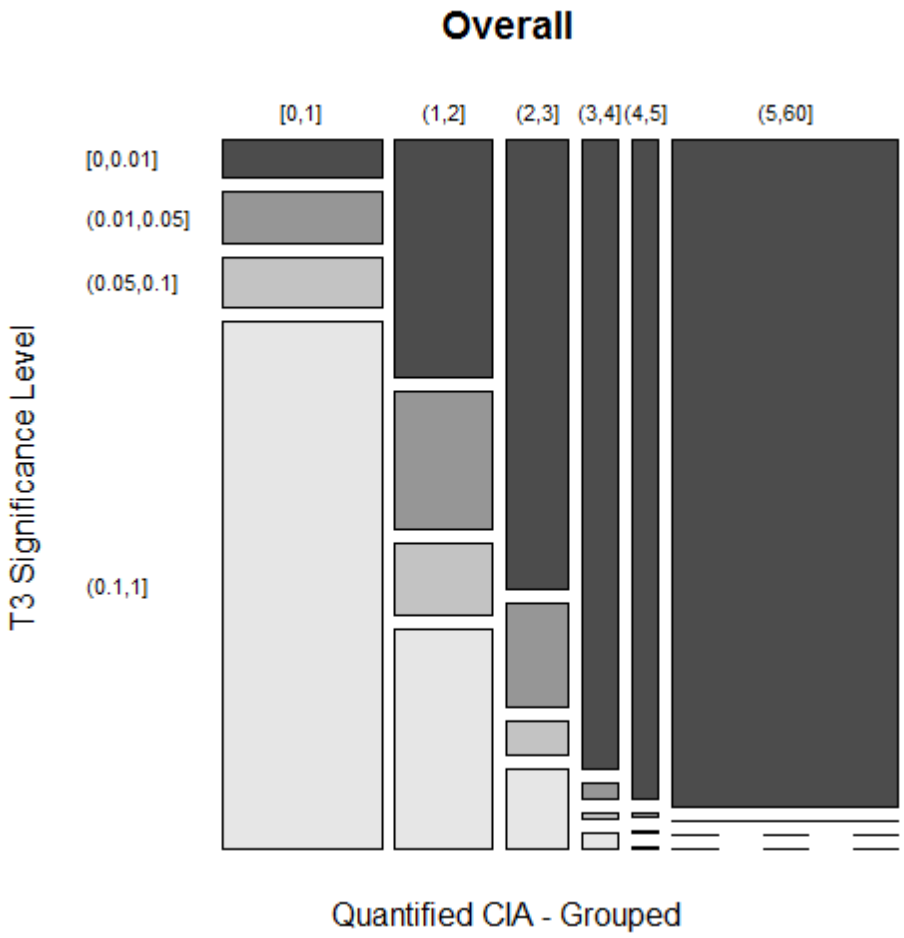

Figure 5.11: Mosaic-plot of quantified CIA by $\tilde{T}3$ p-values

The mosaic-plot is constructed from the frequencies shown in Table 5.19. Row and column percentages for this contingency tables are given in Tables 5.20 and 5.21. The mosaic-plot for all 30,000 simulations (Figure 5.11) gives a strong indication that any deviation from conditional independence has a negative effect on the quality of the fusion in terms of the correlation structure. The size of the grey areas in this plot relative to the black areas changes drastically with deviations from CIA.

The value of the qCIA is low in 75.8% of the simulations for which the correlation structure was effectively retained in the fused data, i.e. p-values greater than 0.1 (Table 5.20). Also, when the fusion is not successful and the test is significant at the 1% level, there is a high probability that conditional independence is not present in the data. The value of the qCIA is low for only 2.4% of these simulations. A large percentage of simulations that were significant between 1-5% and 5-10% also have relatively low qCIA values.

Since the lowest category of the qCIA measure ranges from zero to one, it appears that even very small deviations from the assumption will lead to incorrect results. This constitutes approximately 20% of all simulations within the lowest qCIA category (Table 5.21). For a qCIA value as low as between 1 and 2, two thirds of the simulations will not exhibit the same correlation structure in the fused file as in the original data. For any qCIA greater than 2 the fusion is unable to accurately reflect the true correlation structure.

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	442	1705	2080	1670	1312	11133	18342
(0.01, 0.05]	626	993	485	45	9	2	2160
(0.05, 0.1]	578	521	159	14	3	0	1275
(0.1, 1]	6236	1574	367	43	3	0	8223
TOTAL	7882	4793	3091	1772	1327	11135	30000

Table 5.19: Frequencies of categorized qCIA by $\tilde{T}3$ p-values category

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	2.4	9.3	11.3	9.1	7.2	60.7	100
(0.01, 0.05]	29.0	46.0	22.5	2.1	0.4	0.1	100
(0.05, 0.1]	45.3	40.9	12.5	1.1	0.2	0.0	100
(0.1, 1]	75.8	19.1	4.5	0.5	0.0	0.0	100
TOTAL	26.3	16.0	10.3	5.9	4.4	37.1	100

Table 5.20: Row % of categorized qCIA by $\tilde{T}3$ p-values category

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	5.6	35.6	67.3	94.2	98.9	100.0	61.1
(0.01, 0.05]	7.9	20.7	15.7	2.5	0.7	0.0	7.2
(0.05, 0.1]	7.3	10.9	5.1	0.8	0.2	0.0	4.3
(0.1, 1]	79.1	32.8	11.9	2.4	0.2	0.0	27.4
TOTAL	100	100	100	100	100	100	100

Table 5.21: Column % of categorized qCIA by $\tilde{T}3$ p-values category

Level 2: Preserving joint distributions

The most important test of a fusion success is the evaluation of the joint distribution of the variables that were never jointly observed. Figure 5.12 shows the relationship between the levels of CIA and the results from the Chi-squared test, i.e. the p-values, for the joint distribution between the unique variables. It is clear that there is not a gradual decline when deviating from conditional independence. The second graph (Figure 5.13) is focused on a portion of the scatter-plot, for qCIA values up 5.

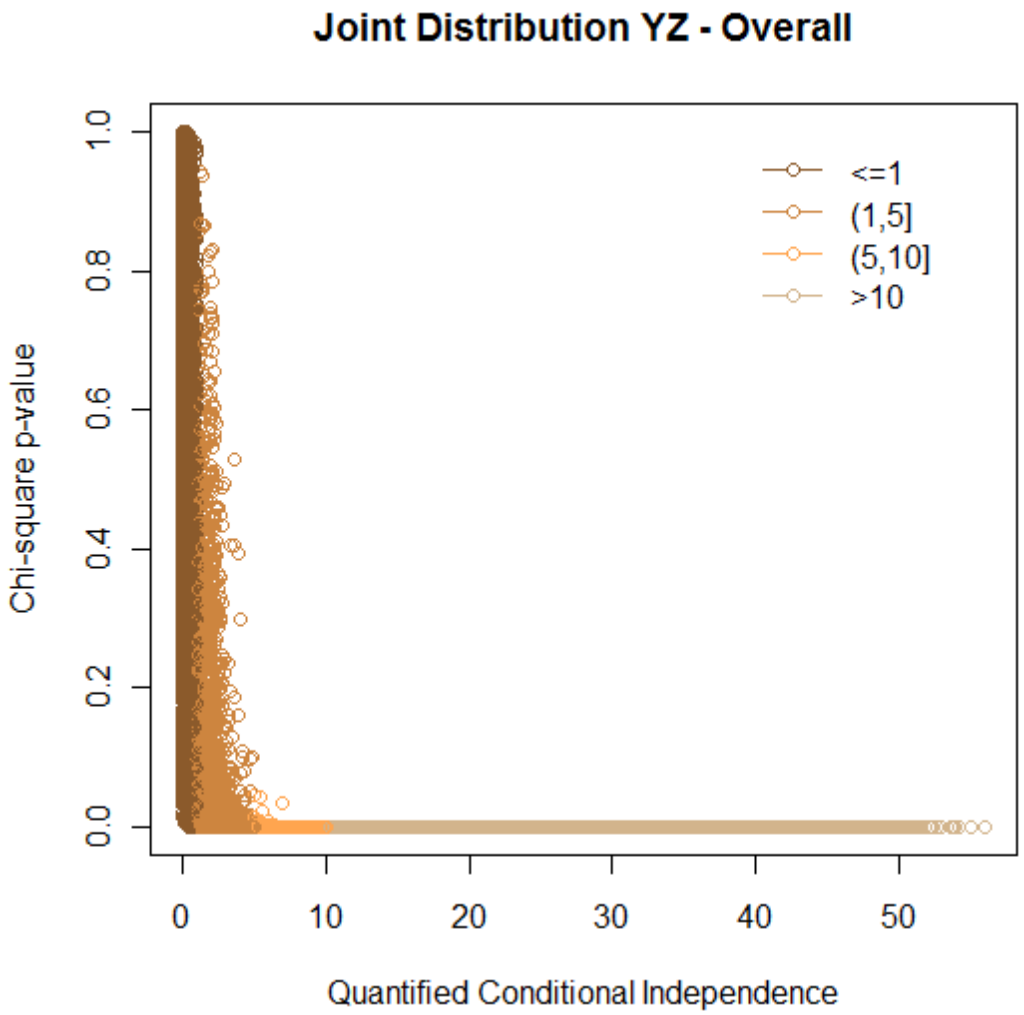

Figure 5.12: Scatter-plot of quantified CIA by χ^2 p-values for (Y, **Z**)

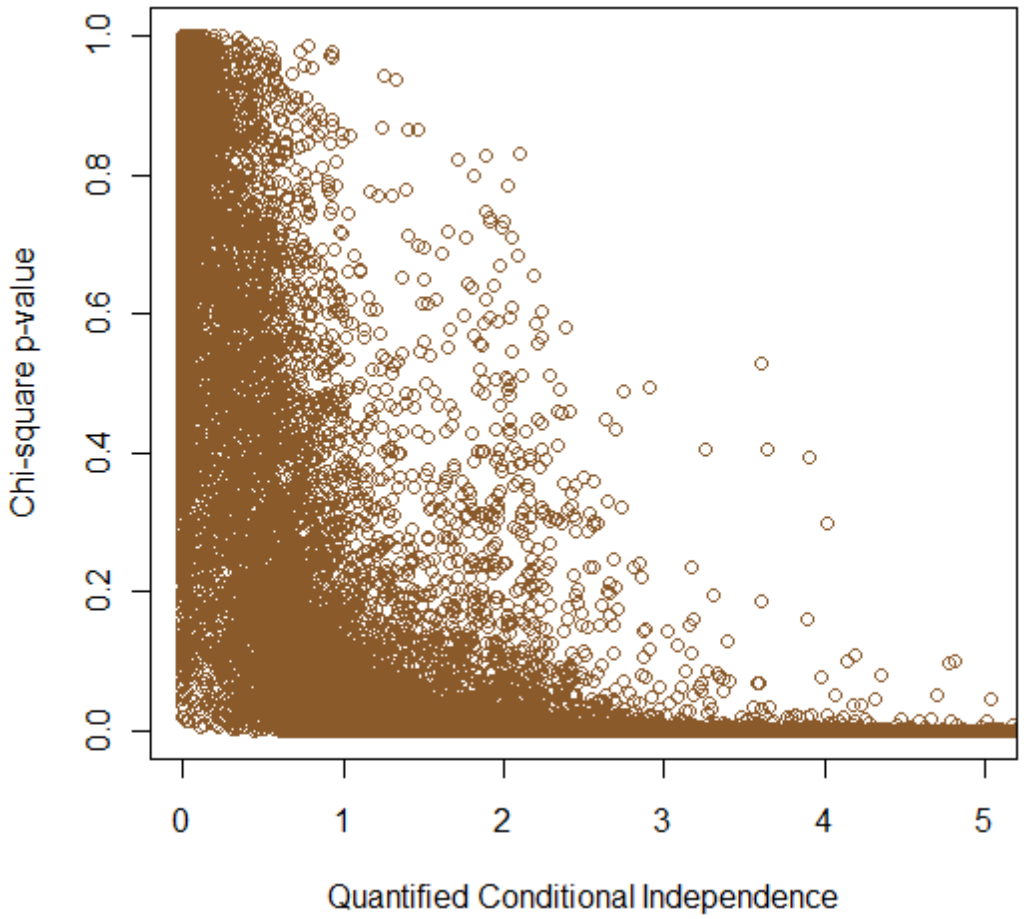

Figure 5.13: Scatter-plot of quantified CIA (≤ 5) by χ^2 p-values for (Y, **Z**)

These two figures show that the fusion deteriorates abruptly as the data deviate from conditional independence. This pattern is better represented using a mosaic-plot. The frequencies used to create the mosaic-plot (Figure 5.14), as well as the row and column percentages for this contingency table are given in Tables 5.22 to 5.24.

These results are very similar to the results from the $\tilde{T}3$ test. If the joint distribution was retained in the fused data, the qCIA value is close to zero, while significant differences between the fused and original distribution become apparent when the assumption of conditional independence is no longer valid (Table 5.23).

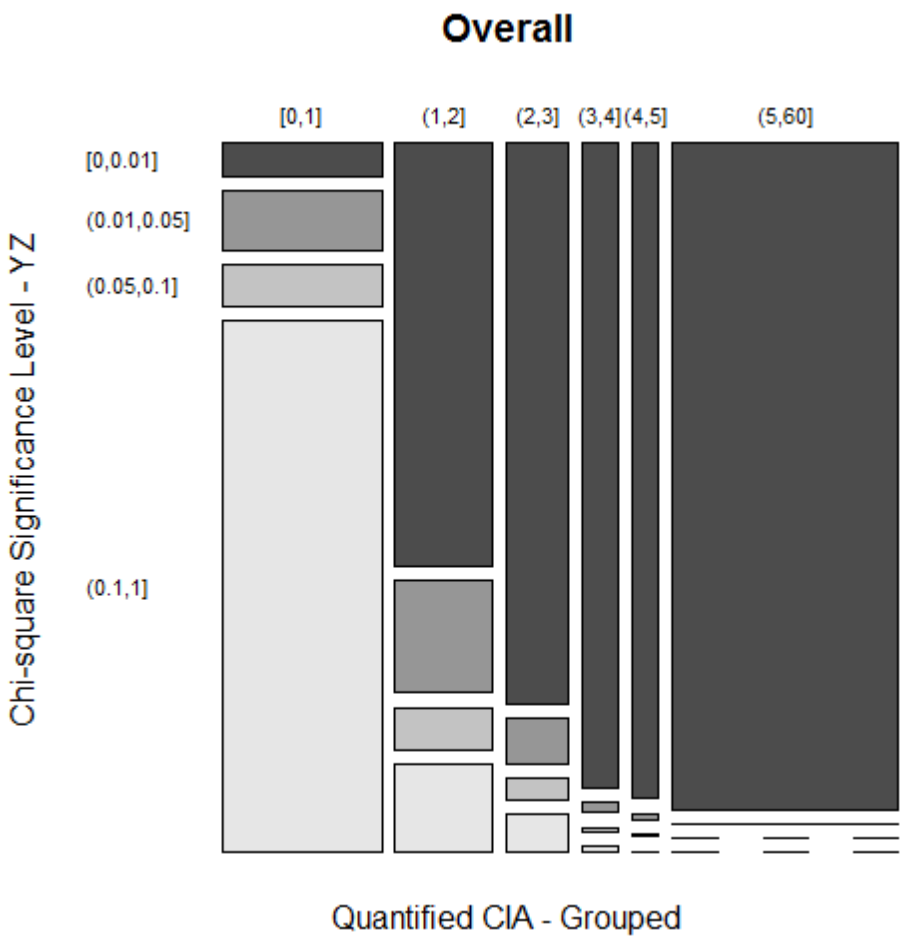

Figure 5.14: Mosaic-plot of categorized qCIA by χ^2 p-values for (Y, **Z**)

Table 5.24 shows that when the qCIA measure is less than or equal to one, 79.6% of simulations reflect the true joint distribution of Y and **Z**. Even for very small deviations from conditional independence, such as the interval (1, 2], the picture changes dramatically and only 13.1% of the simulations accurately fused the unique variables.

These results indicate that a fusion can only truly be successful if there is complete conditional independence in the data.

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	393	3048	2602	1714	1306	11130	20193
(0.01, 0.05]	703	805	212	29	13	5	1767
(0.05, 0.1]	514	312	97	15	6	0	944
(0.1, 1]	6272	628	180	14	2	0	7096
TOTAL	7882	4793	3091	1772	1327	11135	30000

Table 5.22: Frequencies of categorized qCIA by χ^2 p-value categories for (Y, **Z**)

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	1.9	15.1	12.9	8.5	6.5	55.1	100
(0.01, 0.05]	39.8	45.6	12.0	1.6	0.7	0.3	100
(0.05, 0.1]	54.4	33.1	10.3	1.6	0.6	0.0	100
(0.1, 1]	88.4	8.9	2.5	0.2	0.0	0.0	100
TOTAL	26.3	16.0	10.3	5.9	4.4	37.1	100

Table 5.23: Row % of categorized qCIA by χ^2 p-value categories for (Y, **Z**)

	[0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 60]	TOTAL
[0, 0.01]	5.0	63.6	84.2	96.7	98.4	100.0	67.3
(0.01, 0.05]	8.9	16.8	6.9	1.6	1.0	0.0	5.9
(0.05, 0.1]	6.5	6.5	3.1	0.8	0.5	0.0	3.1
(0.1, 1]	79.6	13.1	5.8	0.8	0.2	0.0	23.7
TOTAL	100	100	100	100	100	100	100

Table 5.24: Column % of categorized qCIA by χ^2 p-value categories for (Y, **Z**)

Level 1: Preserving individual values

The hit rate of a fusion shows the proportion of original records that were recreated or retained in the data fusion. It is expected that if the assumption of conditional independence is true in the data, most of the records will be retained. Figure 5.15 shows the relationship between the hit rate for all simulations and the qCIA. The negative linear trend in this graph suggests that any deviation from conditional independence leads to a reduction in the hit rate of the fusion.

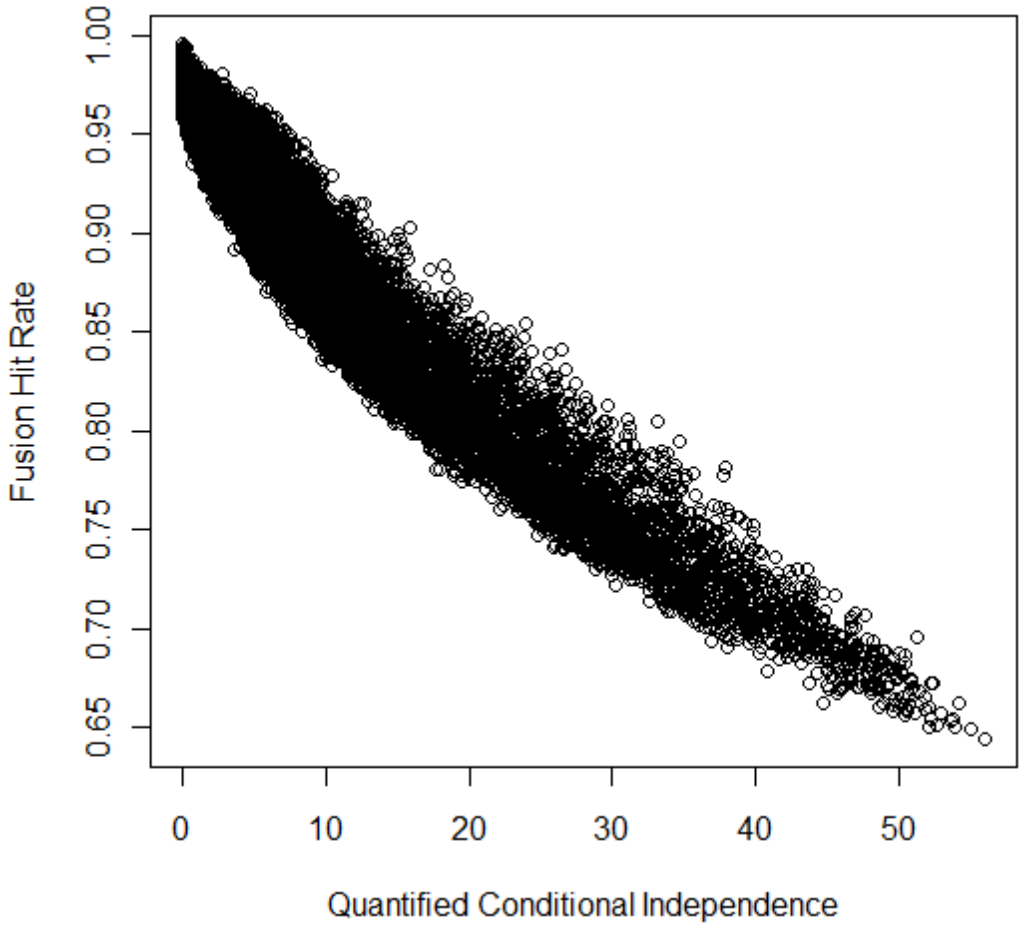

Figure 5.15: Scatter-plot of fusion hit rate

5.3 Analysis by Strength of Correlation

A further research question to investigate is whether the strength of the relationship between the variables Y and **Z** influences the success of the fusion. Since the minimum requirement for a successful fusion is satisfied for all simulations (Table 5.16), only three levels of validity will be assessed in this section, namely the correlation structure evaluation ($\tilde{T}3$), the Chi-squared goodness-of-fit test for the joint distribution of the unique variables, and the hit rate of the fusion. This is done for the two different levels of correlation strength (strong and weak).

Level 3: Preserving correlation structures

Figures 5.16 shows the percentage of simulations within the categorized levels of conditional independence that were not significant for the $\tilde{T}3$ tests performed on data with strong and weak correlations between the unique variables. The number of simulations per qCIA category is given in Table 5.25 for both weak and strong correlations.

	[0, 1]	**(1, 2]**	**(2, 3]**	**(3, 4]**	**(4, 5]**	**(5, 60]**	**TOTAL**
Weak	7154	4034	2081	709	225	158	14361
Strong	728	758	1009	1061	1102	10904	15562

Table 5.25: Frequencies within qCIA categories for weak and strong correlations

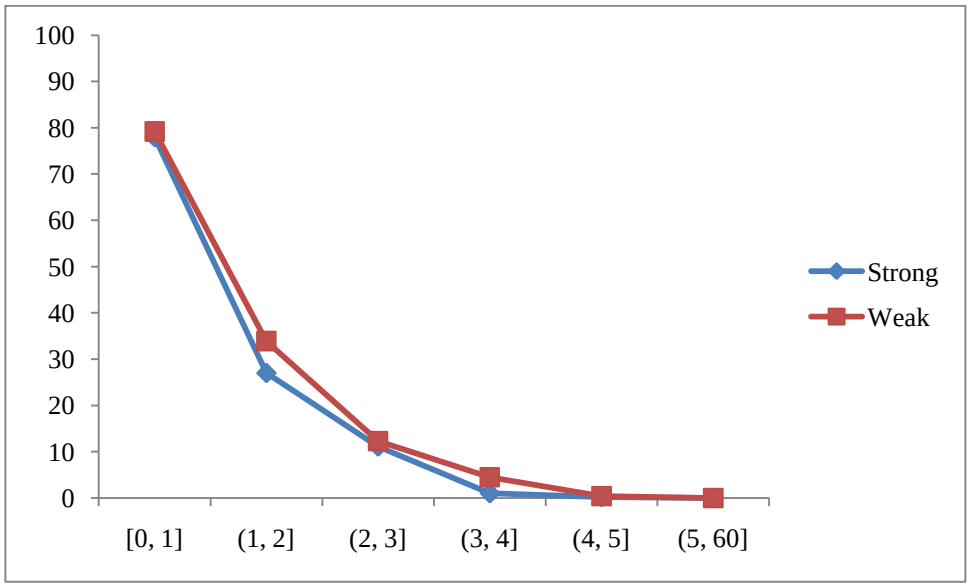

Figure 5.16: Percentage non-significant $\tilde{T}3$ p-values within qCIA categories

Approximately 80% of simulations for which the qCIA measure is less than or equal to one (Figure 5.16) are correctly reproduced in the fusion, for both strong and weak correlation structures. The fusion deteriorates very quickly with respect to the correlation structure for deviations from conditional independence. Approximately two thirds of the simulations produce fused distributions that are significantly different from the original distribution. This trend is the same regardless of the level of correlation between the unique variables.

Level 2: Preserving joint distributions

The difference between data with weak vs. strong correlations is more pronounced for the Chi-squared test for the joint distribution (Y, **Z**), given in Figure 5.17. For qCIA values in the interval (1, 2], approximately 10% of the simulations with a weak correlation structure are not significant, compared to more than 30% for simulations with a strong correlation structure. This implies that the test is more likely to be significant for data with weak correlations than for data with strong correlations. Despite these differences, there is a considerable decline in the success of the fusion

with deviations from conditional independence, for data with both weak and strong correlation structures.

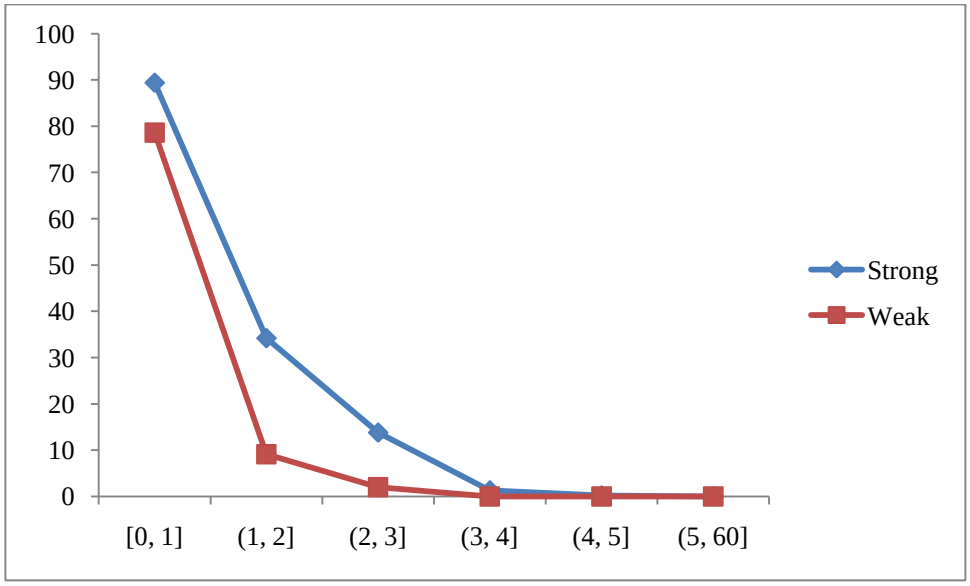

Figure 5.17: Percentage non-significant χ^2 p-values for (Y, **Z**) within qCIA categories

Level 1: Preserving individual values

The hit rate for strong vs. weak correlations follow the same trend (Figure 5.18). As the data deviate from conditional independence, the hit rate reduces, and consequently the success of the fusion deteriorates.

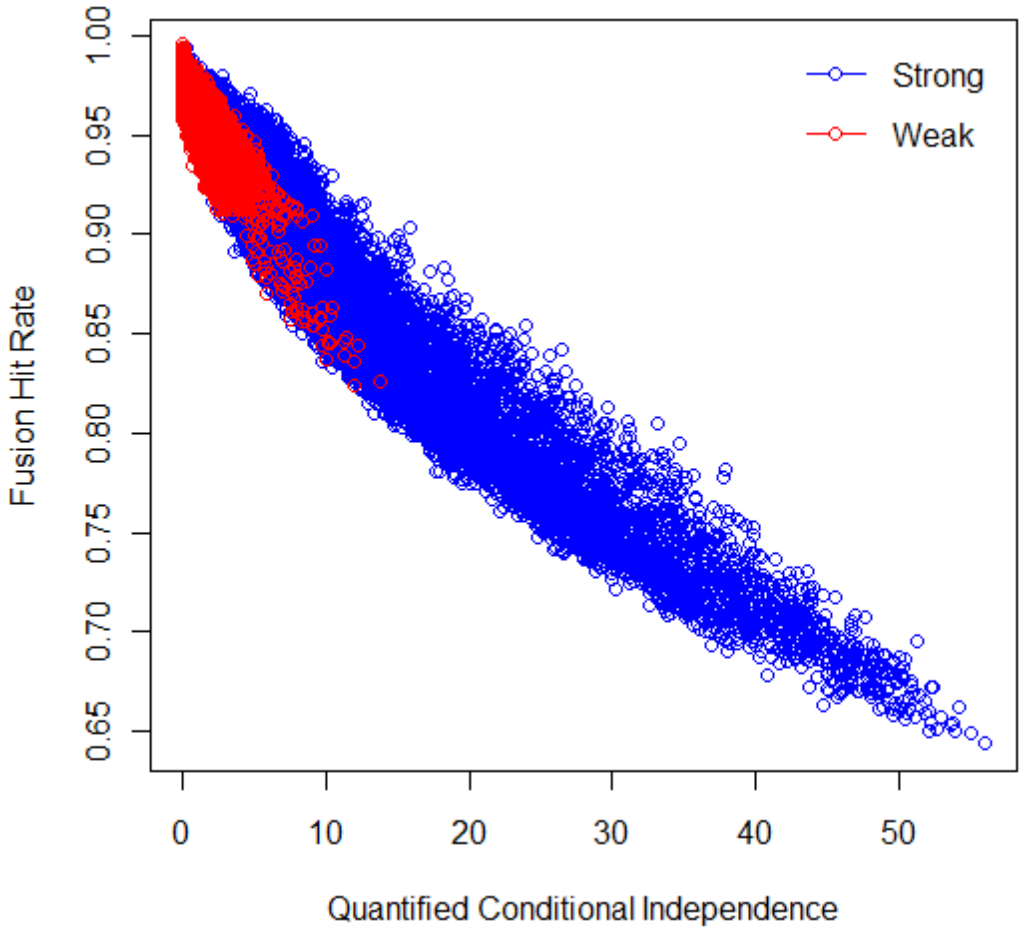

Figure 5.18: Scatter-plot of fusion hit rate for two levels of correlation

5.4 Analysis for Partial CIA

In practical situations, it is possible that the CIA is valid for a subset of the unique variables, but not for all. This is referred to as partial conditional independence. To assess the impact of this situation on the quality of the fusion, the simulations for which this is true must first be identified. Initial inspection of the relationship between the two partial correlations $\rho(YZ_1 \mid X)$ and $\rho(YZ_2 \mid X)$, shows that this situation does occur for some simulations (Figure 5.19).

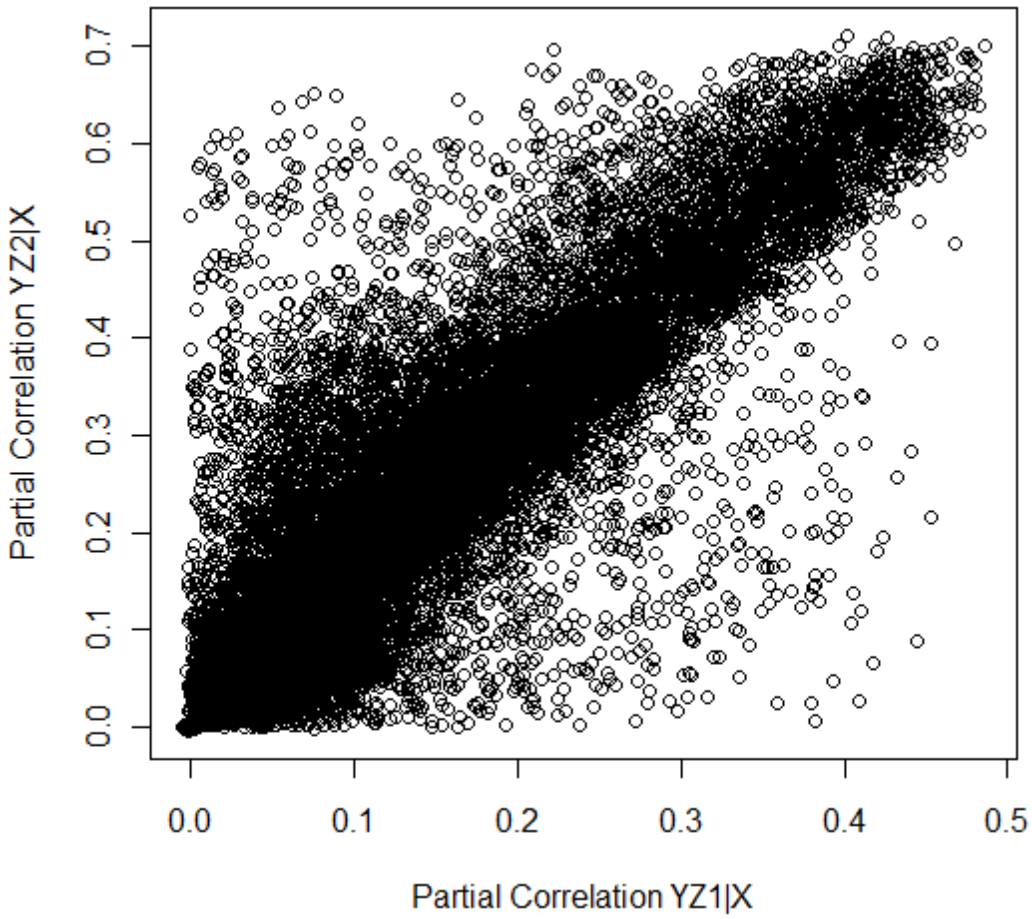

Figure 5.19: Scatter-plot of partial correlations

It was shown in equation (4.3) that, for a sample size of 2000, a partial correlation of 0.05 is significantly different from zero, indicating the absence of conditional independence. Two separate groups that exhibit partial conditional independence are identified based on this result. Group 1 consists of all simulations for which CIA exists between variables Y and Z_1, but not between Y and Z_2. For this group the values of the partial correlations are $\rho(YZ_1 \mid X) < 0.05$ and $\rho(YZ_2 \mid X) \geq 0.05$. The second group shows the converse, namely $\rho(YZ_1 \mid X) \geq 0.05$ and $\rho(YZ_2 \mid X) < 0.05$. The number of simulations in each group is n_1 = 3520 and n_2 = 870 respectively.

The average qCIA measure for all 30,000 simulations is 6.67. Although the qCIA values within each group could be large, the average value for both Group 1 and Group 2 is generally much lower than the overall average (Table 5.26).

	Quantified Conditional Independence	
	Average	Range
Group 1	2.13	(0.21, 31.50)
Group 2	1.53	(0.33, 14.42)

Table 5.26: Average qCIA values for partial CIA groups, with ranges

Figures 5.20 and 5.21 show the percentage of simulations for which the fused distributions were significantly different from the original distributions, for Groups 1 and 2 respectively. This is given for the three joint distributions (Y, Z), (Y, Z_1) and (Y, Z_2). The Chi-squared p-values are categorized into three levels: up to 5%, between 5-10%, and not significant at the 10% level.

If the CIA is valid for the relationship between Y and Z_1 only, the distribution of these variables is retained in 84% of the simulations (Figure 5.20). However, the (Y, Z_2) distribution is not preserved. As a result, the overall fusion is not successful. Similarly, the joint distribution of (Y, Z_2) for Group 2 is effectively estimated through the fusion, but the overall distribution is not retained (Figure 5.21).

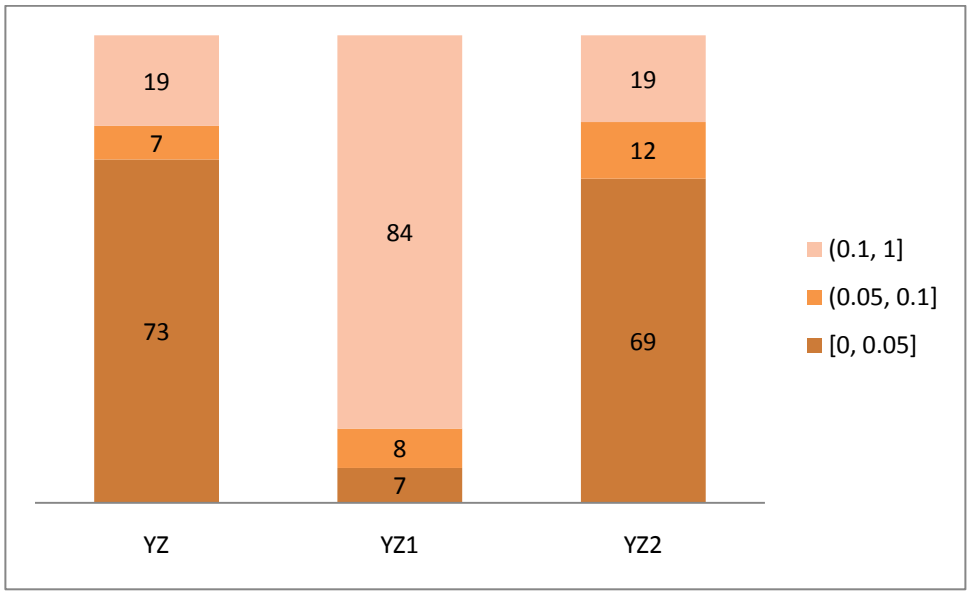

Figure 5.20: Bar-chart of χ^2 p-value categories for Group 1

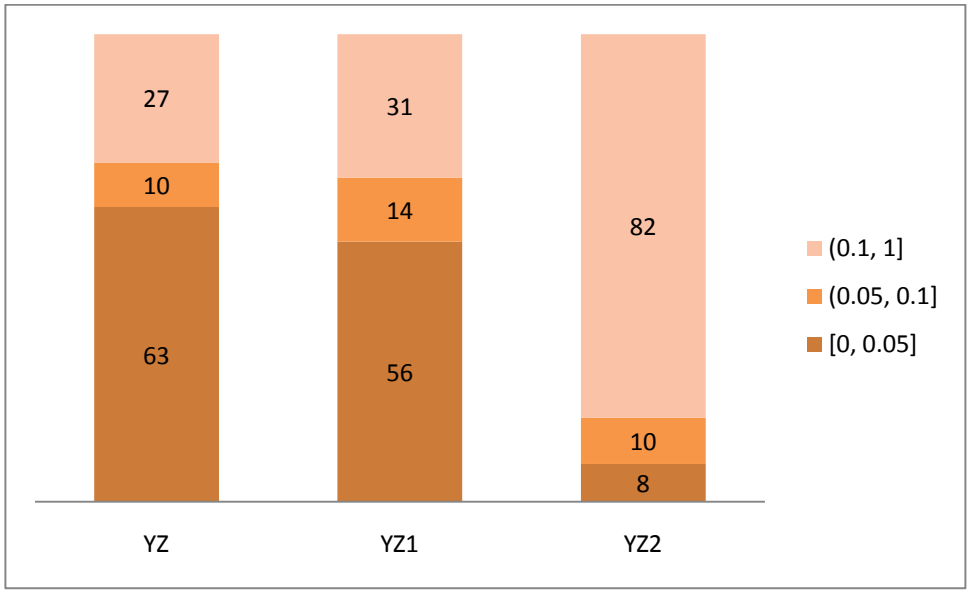

Figure 5.21: Bar-chart of χ^2 p-value categories for Group 2

5.5 Summary

The analyses in sections 5.2 to 5.4 specifically address the research questions posed in Chapter 1. A key objective of this analysis is to determine the success of binary data fusion for different levels of CIA in the data. In order to assess whether deviations from CIA will influence the success of binary data fusion, it is important to be able to first quantify the degree of CIA present in the data.

Section 5.2.2 evaluates the feasibility of the qCIA measure, derived from the CMI, to quantify conditional independence. This measure is zero or close to zero when both partial correlations, $\rho(YZ_1 \mid X)$ and $\rho(YZ_2 \mid X)$, are close to zero. Simulations where both partial correlations are approximately equal to zero imply that the CIA is a valid assumption. As the data deviate from the CIA, at least one of the partial correlations is non-zero. For such simulations, the qCIA measure also deviates from zero. Therefore, the qCIA measure effectively captures the degree of the CIA present in the binary data.

The results in section 5.2.4 (fusion evaluation) show that binary data fusion will be successful if, and only if, the CIA is true. This is evident from the hit rate of the fusion, the $\tilde{T}3$ test for a correlation structure, as well as the Chi-squared goodness-of-fit test for the joint distribution of the unique variables (Y, **Z**). The scatter-plot of the fusion hit rate shows that the proportion of original records retained in the fusion decreases with deviations from CIA. The comparison between p-values and the qCIA measure for both statistical tests ($\tilde{T}3$ and χ^2) indicates that the success of the fusion is immediately compromised if the CIA is not valid. The success of a fusion of data with small deviations from the assumption cannot be guaranteed.

The success of a binary fusion does not depend on the strength of the relationship between the unique variables, as shown in section 5.3. The hit rate trend is the same for simulations with both weak and strong correlation structures. The proportion of

non-significant $\tilde{T}3$ p-values within qCIA categories, for weak vs. strong correlations, are very similar and indicate that the correlation structure is only retained in the fusion if the CIA is valid. The Chi-squared test for the joint distribution (Y, **Z**) shows that the success of a fusion deteriorates quicker if the data have a weak correlation structure compared to data with a strong correlation structure. However, the overall interpretation remains the same, namely that the CIA must be true for the fusion to produce valid results, regardless of the strength of the relationship between Y and **Z**.

In the situation that the CIA is valid for only a subset of the unique variables, the fusion will only produce accurate estimates for the variables for which the CIA is true, and not for the other variables (section 5.4). This impacts on the overall success of the fusion in that it is not possible to accurately estimate the entire joint distribution of all the unique variables. Although partial CIA can occur in data, there is not sufficient information available in practical fusion applications to identify the subset of variables for which CIA is true. In order to successfully use data fusion, the CIA must be valid of all the unique variables.

6 CONCLUSIONS

6.1 Conclusions

The validity of the CIA in data fusion has raised much concern in the statistical literature. It has been shown by numerous authors that the assumption is perhaps too restrictive to be considered a reliable data fusion methodology. A major drawback of the CIA is that, in practical situations, it is not possible to test whether the assumption is valid or not. Therefore, there is the risk of fusing data based on an incorrect assumption, which leads to the fused data set distorting the true distributions of the data.

Although most simulation exercises in the literature are done for continuous variables, the use of simulated binary data produced results that correspond to findings from other simulation applications with respect to the success of a fusion under the assumption of conditional independence. When this assumption is true, the binary data fusion is guaranteed to be a success. All the different distributions and data structures are sufficiently retained in the fused data, thereby satisfying all four levels of validity proposed by Rässler (2002).

The main objective of this report is to assess the impact of deviations from the CIA on the quality of a binary fusion. In order to do this, the degree of conditional independence in the data that will reflect deviations from the assumption must be quantified. The CMI measure was used to quantify the degree of conditional independence for each simulated data set.

Using this measure, the results show that deviations from the CIA have a negative effect on the success of a fusion. Even small deviations did not always produce

accurate results. Regardless of the level of correlation between the unique variables, the fusion deteriorates very quickly for deviations from conditional independence.

In the event that the assumption is valid for only a subset of the unique variables, the fusion will not accurately reflect the overall distribution. Only those distributions for which the CIA is true will be correctly reproduced in the fused file. The problem is that, in practical fusion applications, it is impossible to determine which joint distributions will be true in the fused file and which will be incorrect, since no joint information is available to evaluate this.

This really raises the question: how much confidence can the researcher truly have in the validity of a data fusion under the assumption of conditional independence? Although fusion may be the only viable solution to the problem of questionnaire overload, it can only be done if there is sufficient evidence that the required assumptions are satisfied. In market research applications, fusion done on data for which the CIA is not valid will lead to a misrepresentation of the attitudes and behaviour in the market. Any conclusions drawn from such fused data will be seriously flawed.

6.2 Recommendations

Research into data fusion in general, as well as fusing binary data, is far from complete. This analysis has shown that binary data fusion will be successful if the CIA is a valid assumption, but alternative approaches that do not rely on this assumption should be investigated. In recent years, multiple imputation has been on the forefront of fusion research, with the main focus on continuous variables. No literature currently exists for multiple imputation techniques to fuse binary data. This area of research presents important opportunities for further study, since it not only investigates ways of fusing categorical data, but can also contribute to the area of missing data analysis when data are categorical.

Any researcher attempting data fusion must ensure that the CIA is valid for the data. This is very difficult to verify, particularly for ad-hoc fusion where the data were collected independently through different sources. In such cases it is important that the data are adequately aligned in terms of the common variables and the response units. Paass (1986) first suggested the use of auxiliary data to verify the validity of the CIA and to improve the model. Without such external information it is unrealistic to assume that the assumption is satisfied in the data. The time that the auxiliary data were collected, the target population and the structure of the common variables must also be thoroughly inspected to ensure that the data are usable.

Planned fusion may produce better fusion results under the assumption of conditional independence if the division of the larger survey into parts is done effectively. It is vital that the researcher correctly identifies the variables for which this assumption is valid. The quantified conditional independence measure can potentially aid in this. This is an area for future research, and the properties of this measure could be further investigated in terms of how and when it can be used, and whether it is able to identify the subset of variables in a questionnaire for which the CIA is a valid assumption.

7 REFERENCES

Alosh, M.A. and Lee, S. J. (2001). *A simple approach for generating correlated binary variates*. Journal of Statistical Computation and Simulation, **70**, 231-255.

Alter, H.E. (1974). *Creation of a synthetic data set by linking records of the Canadian survey of consumer finances with the family expenditure survey 1970*. Annals of Economic and Social Measurement **3(**2**),** 373-394.

ARF Guidelines for Data Integration (2003). www.thearf.org/assets/research-standards

Bahadur, R.R. (1961). *A representation of the joint distribution of responses to n dichotomous items*. In Studies in item analysis and prediction. Ed. H. Solomon. Stanford: Stanford University Press, pp. 158-168.

Becker, R. and Collins, J. (2007). *Toward total audience – Integrating magazines' hardcopy and internet site audiences using dynamic segmentation fusion*. Presented to the Advertising Research Foundation. www.mediamark.com/PDF/WP%20Toward%20Total%20Audience.pdf

Budd, E.C. (1972). *Comments*. Annals of Economic and Social Measurement **1**(3), 349-354.

Conti, P.L., Marella, D. and Scanu, M. (2008). *Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators*. Computational Statistics and Data Analysis **53**, 354-365.

Cover, T.M. and Thomas, J.A. (1991). *Entropy, relative entropy and mutual information*. Elements of Information Theory. John Wiley & Sons, Inc, pp.12-49.

Curtin, R., Presser, S. and Singer, E. (2005). *Changes in telephone survey nonresponse over the past quarter century*. Public Opinion Quarterly, **69**(1), 87-98.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical matching: Theory and practice*. England: Wiley, 256 pages.

Dawson-Saunders, B. and Trapp, R.G. (1994). *Basic and clinical biostatistics*. 2nd edn. Connecticut: Appleton & Lange, 344 pages.

De Heer, W. (1999). *International response trends: Results of an international survey*. Journal of Official Statistics **15**(2)**,** 129-142.

Emrich, L.J. and Piedmonte, M.R. (1991). *A method for generating high-dimensional multivariate binary variates*. American Statistician, **45**, 302-304.

Farrell, P.J. and Rodgers-Stewart, K. (2008). *Methods for generating longitudinally correlated binary data*. International Statistical Review, **76**(1), 28-38.

Farrell, P.J. and Sutradhar, B.C. (2006). *A non-linear conditional probability model for generating correlated binary data*. Statistics and Probability Letters, **76**, 353-361.

Galpin, J.S. and Neethling, A. (2004). *Weighting of surveys and data fusion*. Workshop on Official Statistics, presented at the 2004 conference of the South African Statistical Association, Bloemfontein.

Gange, S.J. (1995). *Generating multivariate categorical variates using the iterative proportional fitting algorithm*. American Statistician, **45**, 134-138.

Gilula, Z., McCulloch, R.E. and Rossi, P.E. (2006). *A direct approach to data fusion*. Journal of Marketing Research **43,** 73-83.

Groves, R.M., Cialdini, R.B. and Couper M.P. (1992). *Understanding the decision to participate in a survey*. Public Opinion Quarterly **56,** 475-495.

Groves, R.M., Fowler Jr., F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*. 2nd edn. New Jersey: Wiley, 461 pages.

Hansen, K.M. (2007). *The effects of incentives, interview length, and interviewer characteristics on response rates in a CATI-study*. International Journal of Public Opinion Research, **19**(1), 112-121.

Hertzog, T.H., Scheuren, F. and Winkler, W.E. (2010). *Record linkage*. John Wiley & Sons, Inc. WIREs Comp Stat, **2**, 535-543.

Higham, N.J. (1988). *Computing a nearest symmetric positive semidefinite matrix*. Linear Algebra and its Applications **103**, 103-118.

Howell, D.C. (2009). *Treatment of missing data*. www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html

Ingram, D., O'Hare, J., Scheuren, F. and Turek, J. (2000). *Statistical matching: A new validation case study*. Proceedings of the Survey Research Methods Section, American Statistical Association.
www.amstat.org/sections/SRMS/Proceedings/papers/2000_126.pdf

Jakulin, A. and Bratko, I. (2004). *Quantifying and visualizing attribute interactions: An approach based on entropy*.
http://arxiv.org/abs/cs.AI/0308002

Kadane, J.B. (1975). *Statistical problems of merged data files*. Office of Tax Analysis, U.S. Department of the Treasury, Paper 6.
www.ustreas.gov/offices/tax-policy/library/ota6.pdf

Kang, S.H. and Jung, S.H. (2001). *Generating correlated binary with complete specification of the joint distribution*. Biometrical Journal, **43**, 263-269.

Kanter, M. (1975). *Autoregression for discrete processes mod 2*. Journal of Applied Probability, **12**, 371-375.

Kiesl, H. and Rässler, S. (2006). *How valid can data fusion be?* IAB Discussion Paper. No 15/2006.

Larntz, K. and Perlman, M.D. (1985). *A simple test for the equality of correlation matrices*. Technical Report No. 63.
www.stat.washington.edu/research/reports/1985/tr063.pdf

Lee, A.J. (1993). *Generating random binary deviates having fixed marginal distributions and specified degrees of freedom*. American Statistician, **47**, 209-215.

Leisch, F., Weingessel, A. and Hornik, K. (1998). *On the generation of correlated artificial binary data*. Working Paper 13, Institute für Statistik und Wahrscheinlichkeitstheorie, Vienna University of Technology, Austria.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical analysis with missing data*. New Jersey: Wiley, 278 pages.

Lohr, S.L. (1999). *Sampling: Design and analysis*. Duxberry Press, Pacific Grove, 494 pages.

Lunn, A.D. and Davies, S.J. (1998). *A note on generating correlated binary variables*. Biometrika, **85**, 487-490.

McKenzie, E. (1981). *Extending the correlation structure of exponential autoregressive moving average processes*. Journal of Applied Probability, **18**, 1-9.

Moriarity, C. and Scheuren, F. (2004). *Regression-based statistical matching: Recent developments*. Proceedings of the section on survey research methods, American Statistical Association. www.amstat.org/sections/srms/Proceedings/y2004/files/Jsm2004-000361.pdf

O'Brien, S. (1991). *The role of data fusion in actionable media targeting in the 1990's*. Marketing and Research Today, **19**, 15-22.

Okner, B.A. (1972). *Constructing a new data base from existing microdata sets: The 1966 MERGE file*. Annals of Economic and Social Measurement **1**(3)**,** 325-342.

Oman, S.D. and Zucker, D.M. (2001). *Modelling and generating correlated binary variables*. Biometrika, **88**, 287-290.

Paass, G. (1986). *Statistical match: Evaluation of existing procedures and improvements by using additional information*. In Micro-analytic Simulation Models to Support Social and Financial Policy. Ed. Orcutt, Merc and Quinke. Elsevier Science, Amsterdam, pp. 401-422.

Park, C.G., Park, T. and Shin, D.W. (1996). *A simple method for generating correlated binary variates*. American Statistician, **50**(4), 306-310.

Qaqish, B.F. (2003). *A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations*. Biometrika, **90**(2), 455-463.

Radner, D. B., Allen, R., Gonzalez, M. E., Jabine, T. B. and Muller, H. J. (1980). *Report on exact and statistical matching techniques*. Statistical Policy Working Paper 5, Federal Committee on Statistical Methodology.

Raghunathan, T.E. and Grizzle, J.E. (1995). *A split questionnaire survey design*. Journal of the American Statistical Association, **90**, 54-63.

Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Lecture Notes in Statistics, 168. New York: Springer, 238 pages.

Rässler, S. and Fleischer, K. (1998). *Aspects concerning data fusion techniques*. ZUMA Nachrichten Spezial **4**, 317-333.

Redway, H. (2003). *Data fusion by statistical matching*. https://guard.canberra.edu.au/natsem/conference2003/papers/pdf/redway_howard-1.pdf

Rodgers, W.L. (1984). *An evaluation of statistical matching*. Journal of Business and Economic Statistics, **2**(1), 91-102.

Rubin, D.B. (1977). *The design of a general and flexible system for handling nonresponse in sample surveys*. American Statistician, 2004, **58**(4), 298-302.

Rubin, D.B. (1986). *Statistical matching using file concatenation with adjusted weights and multiple imputations*. Journal of Business and Economic Statistics, **4**(1), 87-94.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New Jersey: Wiley, 258 pages.

Ruggles, N., Ruggles, R. and Wolff, E. (1977). *Merging microdata: Rationale, practice and testing*. Annals of Economic and Social Measurement **6,** 429-444.

Scheuren, F. (2009). Comment made while presenting Moriarity, C. *Regression-based statistical matching: past, present and Future*, during the 57th Session of the ISI, 16-22 August 2009, Durban, South Africa.

Shannon, C.E. (1948). *A mathematical theory of communication*. The Bell System Technical Journal **27,** 379-423, 623-656.

Sims, C.A. (1972). *Comments*. Annals of Economic and Social Measurement **1**, 343-346.

Singer, E. (2006). *Introduction: Nonresponse bias in household surveys*. Public Opinion Quarterly **70**(5)**,** 637-645.

Sivazlian, B.D. and Stanfel, L.E. (1975). *Optimization techniques in operations research*. Prentice-Hall, 502 pages.

Soong, R. and De Montigny, M. (2001). *The anatomy of data fusion*. 2001 Worldwide Readership Research Symposium. www.readershipsymposium.org/anatomy-data-fusion

Tchaoussoglou C. and Van der Noort, W. (1999). *Divide and unite – Splitting the SummoScanner and data fusion*. Worldwide Readership Research Symposium. www.readershipsymposium.org/divide-and-unite-splitting-summoscanner-and-data-fusion

Van der Noort, W. and Tchaoussoglou, C. (1995). *The importance of being constant*. Worldwide Readership Research Symposium. www.readershipsymposium.org/importance-being-constant-effects-questionnaire-overload

Van der Putten, P., Kok, J.N. and Gupta, A. (2002). *Data fusion through statistical matching*. Paper 185, Center for eBusiness@MIT. http://ebusiness.mit.edu

Weatherburn, C.E. (1952). *A first course in mathematical statistics*. 2nd edn. Cambridge University Press, 271 pages.

Wolff, E. (1977). *Estimates of the 1969 size distribution of household wealth in the US from a synthetic database*. Conference on research in income and wealth. www.nber.org/chapters/c7448.pdf

APPENDIX A: BINARY SIMULATION

INPUT

$$P = (p_X, p_Y, p_{Z_1}, p_{Z_2})' = (0.7, 0.6, 0.8, 0.5)'$$

$$R = \{r_{ij}\} = \begin{pmatrix} 1 & 0.698274 & 0.656700 & 0.483163 \\ & 1 & 0.458556 & 0.337380 \\ & & 1 & 0.430799 \\ & & & 1 \end{pmatrix}$$

ALGORITHM

ITERATION 1

Create initial S-matrix with elements α_{ij}, where $\alpha_{ij} = \dfrac{p_i p_j}{\rho_{ij}\sqrt{p_i q_i p_j q_j} + p_i p_j}$

$$S_1 = \begin{pmatrix} 0.7 & 0.728203 & 0.823075 & 0.759703 \\ & 0.6 & 0.842315 & 0.784025 \\ & & 0.8 & 0.822775 \\ & & & 0.5 \end{pmatrix}$$

$$T_1 = \{\alpha_{ij}^1 : 0 < \alpha_{ij}^1 < 1; i, j = 1, \ldots, 4\}$$

$$\beta_1 = \max\{\alpha_{ij}^1 : \alpha_{ij}^1 \in T_1\} = 0.842315$$

$$\{r, s\}_1 = \text{index } (i, j) \text{ of } \beta_1 = \{2, 3\}$$

$$A_1 = \{r, s, i, j : 0 < \alpha_{ij}^1 < 1; \alpha_{ij}^1 \in S_1\} = \{1, 2, 3, 4\}$$

ITERATION 2

Update S-matrix: $\alpha_{ij}^2 = \alpha_{ij}^1 / \beta_1 \quad , \forall \{i, j\} \in A_1$

$$S_2 = \begin{pmatrix} 0.831043 & 0.864526 & 0.977159 & 0.901923 \\ & 0.712323 & 1 & 0.930798 \\ & & 0.949764 & 0.976802 \\ & & & 0.593602 \end{pmatrix}$$

$T_2 = \{\alpha_{ij}^2 : 0 < \alpha_{ij}^2 < 1; i, j = 1, \ldots, 4\}$

$\beta_2 = \max\{\alpha_{ij}^2 : \alpha_{ij}^2 \in T_2\} = 0.977159$

$\{r, s\}_2 = \text{index } (i, j) \text{ of } \beta_2 = \{1, 3\}$

$A_2 = \{r, s, i, j : 0 < \alpha_{ij}^2 < 1; \alpha_{ij}^2 \in S_2\} = \{1, 3, 4\}$

ITERATION 3

Update S-matrix: $\alpha_{ij}^3 = \alpha_{ij}^2 / \beta_2 \quad , \forall \{i, j\} \in A_2$

$$S_3 = \begin{pmatrix} 0.850469 & 0.864526 & 1 & 0.923005 \\ & 0.712323 & 1 & 0.930798 \\ & & 0.971964 & 0.999635 \\ & & & 0.607478 \end{pmatrix}$$

$T_3 = \{\alpha_{ij}^3 : 0 < \alpha_{ij}^3 < 1; i, j = 1, \ldots, 4\}$

$\beta_3 = \max\{\alpha_{ij}^3 : \alpha_{ij}^3 \in T_3\} = 0.999635$

$\{r, s\}_3 = \text{index } (i, j) \text{ of } \beta_3 = \{3, 4\}$

$A_3 = \{r, s, i, j : 0 < \alpha_{ij}^3 < 1; \alpha_{ij}^3 \in S_3\} = \{3, 4\}$

ITERATION 4

Update S-matrix: $\alpha_{ij}^4 = \alpha_{ij}^3 / \beta_3 \quad , \forall \{i, j\} \in A_3$

$$S_4 = \begin{pmatrix} 0.850469 & 0.864526 & 1 & 0.923005 \\ & 0.712323 & 1 & 0.930798 \\ & & 0.972320 & 1 \\ & & & 0.607700 \end{pmatrix}$$

$T_4 = \{\alpha_{ij}^4 : 0 < \alpha_{ij}^4 < 1; i, j = 1, \ldots, 4\}$

$\beta_4 = \max\{\alpha_{ij}^4 : \alpha_{ij}^4 \in T_4\} = 0.972320$

$\{r, s\}_4 = \text{index } (i, j) \text{ of } \beta_4 = \{3, 3\}$

$A_4 = \{r, s, i, j : 0 < \alpha_{ij}^4 < 1; \alpha_{ij}^4 \in S_4\} = \{3\}$

ITERATION 5

Update S-matrix: $\alpha_{ij}^5 = \alpha_{ij}^4 / \beta_4 \quad , \forall \{i, j\} \in A_4$

$$S_5 = \begin{pmatrix} 0.850469 & 0.864526 & 1 & 0.923005 \\ & 0.712323 & 1 & 0.930798 \\ & & 1 & 1 \\ & & & 0.607700 \end{pmatrix}$$

$T_5 = \{\alpha_{ij}^5 : 0 < \alpha_{ij}^5 < 1; i, j = 1, \ldots, 4\}$

$\beta_5 = \max\{\alpha_{ij}^5 : \alpha_{ij}^5 \in T_5\} = 0.930798$

$\{r, s\}_5 = \text{index } (i, j) \text{ of } \beta_5 = \{2, 4\}$

$A_5 = \{r, s, i, j : 0 < \alpha_{ij}^5 < 1; \alpha_{ij}^5 \in S_5\} = \{1, 2, 4\}$

ITERATION 6

Update S-matrix: $\alpha_{ij}^{6} = \alpha_{ij}^{5} / \beta_{5} \quad , \forall \{i, j\} \in A_{5}$

$$S_{6} = \begin{pmatrix} 0.913698 & 0.928801 & 1 & 0.991627 \\ & 0.765282 & 1 & 1 \\ & & 1 & 1 \\ & & & 0.652880 \end{pmatrix}$$

$T_{6} = \{\alpha_{ij}^{6} : 0 < \alpha_{ij}^{6} < 1; i, j = 1, \ldots, 4\}$

$\beta_{6} = \max\{\alpha_{ij}^{6} : \alpha_{ij}^{6} \in T_{6}\} = 0.991627$

$\{r, s\}_{6} = \text{index } (i, j) \text{ of } \beta_{6} = \{1, 4\}$

$A_{6} = \{r, s, i, j : 0 < \alpha_{ij}^{6} < 1; \alpha_{ij}^{6} \in S_{6}\} = \{1, 4\}$

ITERATION 7

Update S-matrix: $\alpha_{ij}^{7} = \alpha_{ij}^{6} / \beta_{6} \quad , \forall \{i, j\} \in A_{6}$

$$S_{7} = \begin{pmatrix} 0.921413 & 0.928801 & 1 & 1 \\ & 0.765282 & 1 & 1 \\ & & 1 & 1 \\ & & & 0.658393 \end{pmatrix}$$

$T_{7} = \{\alpha_{ij}^{7} : 0 < \alpha_{ij}^{7} < 1; i, j = 1, \ldots, 4\}$

$\beta_{7} = \max\{\alpha_{ij}^{7} : \alpha_{ij}^{7} \in T_{7}\} = 0.928801$

$\{r, s\}_{7} = \text{index } (i, j) \text{ of } \beta_{7} = \{1, 2\}$

$A_{7} = \{r, s, i, j : 0 < \alpha_{ij}^{7} < 1; \alpha_{ij}^{7} \in S_{7}\} = \{1, 2\}$

ITERATION 8

Update S-matrix: $\alpha_{ij}^{8} = \alpha_{ij}^{7} / \beta_{7} \quad , \forall \{i, j\} \in A_{7}$

$$S_8 = \begin{pmatrix} 0.992046 & 1 & 1 & 1 \\ & 0.823946 & 1 & 1 \\ & & 1 & 1 \\ & & & 0.658393 \end{pmatrix}$$

$T_8 = \{\alpha_{ij}^{8} : 0 < \alpha_{ij}^{8} < 1; i, j = 1, \ldots, 4\}$

$\beta_8 = \max\{\alpha_{ij}^{8} : \alpha_{ij}^{8} \in T_8\} = 0.992046$

$\{r, s\}_8 = \text{index } (i, j) \text{ of } \beta_8 = \{1,1\}$

$A_8 = \{r, s, i, j : 0 < \alpha_{ij}^{8} < 1; \alpha_{ij}^{8} \in S_8\} = \{1\}$

ITERATION 9

Update S-matrix: $\alpha_{ij}^{9} = \alpha_{ij}^{8} / \beta_{8} \quad , \forall \{i, j\} \in A_{8}$

$$S_9 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 0.823946 & 1 & 1 \\ & & 1 & 1 \\ & & & 0.658393 \end{pmatrix}$$

$T_9 = \{\alpha_{ij}^{9} : 0 < \alpha_{ij}^{9} < 1; i, j = 1, \ldots, 4\}$

$\beta_9 = \max\{\alpha_{ij}^{9} : \alpha_{ij}^{9} \in T_9\} = 0.823946$

$\{r, s\}_9 = \text{index } (i, j) \text{ of } \beta_9 = \{2,2\}$

$A_9 = \{r, s, i, j : 0 < \alpha_{ij}^{9} < 1; \alpha_{ij}^{9} \in S_9\} = \{2\}$

ITERATION 10

Update S-matrix: $\alpha_{ij}^{10} = \alpha_{ij}^{9} / \beta_{9} \quad , \forall \{i, j\} \in A_{9}$

$$S_{10} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ & 1 & 1 & 1 \\ & & 1 & 1 \\ & & & 0.658393 \end{pmatrix}$$

$T_{10} = \{\alpha_{ij}^{10} : 0 < \alpha_{ij}^{10} < 1; i, j = 1, \ldots, 4\}$

$\beta_{10} = \max\{\alpha_{ij}^{10} : \alpha_{ij}^{10} \in T_{10}\} = 0.658393$

$\{r, s\}_{10} = \text{index } (i, j) \text{ of } \beta_{10} = \{4, 4\}$

$A_{10} = \{r, s, i, j : 0 < \alpha_{ij}^{10} < 1; \alpha_{ij}^{10} \in S_{10}\} = \{4\}$

PARAMETER SUMMARY

Iteration *l*	β_l	A_l
1	0.842315	{1,2,3,4}
2	0.977159	{1,3,4}
3	0.999635	{3,4}
4	0.972320	{3}
5	0.930798	{1,2,4}
6	0.991627	{1,4}
7	0.928801	{1,2}
8	0.992046	{1}
9	0.823946	{2}
10	0.658393	{4}

$U_l \sim Bernoulli(\beta_l) \quad , l = 1, \ldots, 10$

$C_d = \{U_l : A_l = d; l = 1, \ldots, 10\}$

Variable *d*	C_d
1	{1,2,5,6,7,8}
2	{1,5,7,9}
3	{1,2,3,4}
4	{1,2,3,5,6,10}

DETERMINE COMPLETE JOINT PROBABILITY DISTRIBUTION

The probability of getting a specific configuration for four binary variables is the product of the probabilities that the ten Bernoulli variables are equal to zero or one. All possible outcomes across ten Bernoulli variables are evaluated to assess the configuration that each would generated based on the four sets C_d. The probabilities for each Bernoulli vector of zeros and ones are calculated as follows:

$$\begin{aligned} P(\{1111\}) &= \prod_{l=1}^{10} P(U_l = 1) \\ &= \beta_1 \times \beta_2 \times \beta_3 \times \beta_4 \times \beta_5 \times \beta_6 \times \beta_7 \times \beta_8 \times \beta_9 \times \beta_{10} \\ &= 0.369090 \end{aligned}$$

$$\begin{aligned} P(\{1110\}) &= \prod_{l=1}^{9} P(U_l = 1) \times P(U_{10} = 0) \\ &= \beta_1 \times \beta_2 \times \beta_3 \times \beta_4 \times \beta_5 \times \beta_6 \times \beta_7 \times \beta_8 \times \beta_9 \times (1 - \beta_{10}) \\ &= 0.191502 \end{aligned}$$

The calculated probabilities of all Bernoulli vectors that generate the same configuration for four binary variables are added together to form the probability of getting that particular configuration. The probability associated with the configuration {0000} is calculated using the complement rule. The following table summarizes the Bernoulli vectors of zeros and ones for each configuration of four binary variables, and the associated probabilities.

Binary outcome	Bernoulli outcomes	Probability
0001	1110110001 / 1110111001 / 1110110101 / 1110110011 / 1110110111	0.001003
0010	1111000000 / 1111100000 / 1111010000 / 1111110000 / 1111001000 / 1111101000 / 1111011000 / 1111111000 / 1111000100 / 1111100100 / 1111010100 / 1111110100 / 1111001100 / 1111101100 / 1111011100 / 1111000010 / 1111100010 / 1111010010 / 1111110010 / 1111001010 / 1111011010 / 1111000110 / 1111100110 / 1111010110 / 1111110110 / 1111001110 / 1111011110 / 1111000001 /	0.075112

	1111100001 / 1111010001 / 1111001001 / 1111101001 / 1111011001 / 1111000101 / 1111100101 / 1111010101 / 1111001101 / 1111101101 / 1111011101 / 1111000011 / 1111100011 / 1111010011 / 1111001011 / 1111011011 / 1111000111 / 1111100111 / 1111010111 / 1111001111 / 1111011111	
0011	1111110001 / 1111111001 / 1111110101 / 1111110011 / 1111110111	0.035246
0100	1000101010 / 1100101010 / 1010101010 / 1110101010 / 1001101010 / 1101101010 / 1011101010 / 1000111010 / 1100111010 / 1010111010 / 1110111010 / 1001111010 / 1101111010 / 1011111010 / 1000101110 / 1100101110 / 1010101110 / 1110101110 / 1001101110 / 1101101110 / 1011101110 / 1000111110 / 1010111110 / 1001111110 / 1011111110 / 1000101011 / 1100101011 / 1010101011 / 1110101011 / 1001101011 / 1101101011 / 1011101011 / 1000111011 / 1100111011 / 1010111011 / 1001111011 / 1101111011 / 1011111011 / 1000101111 / 1100101111 / 1010101111 / 1110101111 / 1001101111 / 1101101111 / 1011101111 / 1000111111 / 1010111111 / 1001111111 / 1011111111	0.013888
0101	1110111011	0.000084
0110	1111101010 / 1111111010 / 1111101110 / 1111101011 / 1111101111	0.006307
0111	1111111011	0.002959
1000	1100111100 / 1110111100 / 1101111100 / 1100111101 / 1101111101	0.001210
1001	1110111101	0.002245
1010	1111111100	0.040919
1011	1111111101	0.078864
1100	1100111110 / 1110111110 / 1101111110 / 1100111111 / 1101111111	0.005662
1101	1110111111	0.010507
1110	1111111110	0.191502
1111	1111111111	0.369090
0000	*Complement rule*	0.165400

APPENDIX B: QUANTIFIED CIA

The quantified conditional independence measure is a function of entropy values for selected marginal and joint distributions from the original data set AB. These distributions are given in Tables B.1 to B.7 and the entropy of the distribution is calculated for each table.

X = 0	**X = 1**
0.300150	0.699850

Table B1: Distribution of X in data set AB

$$H(X) = -\sum_{\forall x} p(x) \log_2 p(x) = 0.881474$$

Y = 0	**Y = 1**
0.399700	0.600300

Table B2: Distribution of Y in data set AB

$$H(Y) = -\sum_{\forall y} p(y) \log_2 p(y) = 0.970775$$

	Z_2 = 0	**Z_2 = 1**
Z_1 = 0	0.186093	0.013570
Z_1 = 1	0.314157	0.486243

Table B3: Distribution of Z in data set AB

$$H(Z) = -\sum_{\forall z} p(z) \log_2 p(z) = 1.565919$$

	Y = 0	**Y = 1**
X = 0	0.276638	0.023512
X = 1	0.123062	0.576788

Table B4: Distribution of XY in data set AB

$$H(YX) = -\sum_{\forall y,x} p(y,x)\log_2 p(y,x) = 1.469941$$

	Z_1 = 0		**Z_1 = 0**	
	Z_2 = 0	**Z_2 = 1**	**Z_2 = 0**	**Z_2 = 1**
X = 0	0.179590	0.001001	0.081541	0.038019
X = 1	0.006503	0.012506	0.232616	0.448224

Table B5: Distribution of XZ in data set AB

$$H(ZX) = -\sum_{\forall z,x} p(z,x)\log_2 p(z,x) = 2.063703$$

		Z_1 = 0		**Z_1 = 1**	
		Z_2 = 0	**Z_2 = 1**	**Z_2 = 0**	**Z_2 = 1**
X = 0	**Y = 0**	0.165583	0.001001	0.075038	0.035018
	Y = 1	0.014007	0	0.006503	0.003002
X = 1	**Y = 0**	0.001001	0.002001	0.041021	0.079040
	Y = 1	0.005503	0.010505	0.191596	0.369185

Table B6: Distribution of XYZ in data set AB

$$H(YZX) = -\sum_{\forall y,z,x} p(y,z,x)\log_2 p(y,z,x) = 2.652017$$

Note: When calculating the entropy of a distribution, zero probabilities are excluded from the calculation.

The CMI measure forms the basis of the qCIA measure and is give as

$$\begin{aligned} I(Y,Z \mid X) &= H(YX) + H(ZX) - H(X) - H(YZX) \\ &= 1.469941 + 2.063703 - 0.881474 - 2.652017 \\ &= 0.000153 \end{aligned}$$

This is then expressed as a percentage of the valid range of the CMI, namely the qCIA measure. It is given as

$$\begin{aligned} qCIA &= \frac{I(Y,Z \mid X)}{\min\{H(Y), H(Z)\}} \times 100\% \\ &= \frac{0.000153}{\min\{0.979775, 1.565919\}} \times 100\% \\ &= 0.015739 \end{aligned}$$